Published in final edited form as:

Prev Sci. 2015 October ; 16(7): 978–986. doi:10.1007/s11121-015-0545-z.

Increasing the Sensitivity of Measures to Change

Carlotta Ching Ting Fok and
University of Alaska Fairbanks

David Henry
University of Illinois at Chicago

Abstract

Little attention is paid in prevention research to the ability of measures to accurately assess change, termed “responsiveness” or “sensitivity to change.” This paper reviews definitions and measures of responsiveness, and suggests five strategies for increasing sensitivity to change, with central focus on prevention research with small samples: (a) Improving understandability and cultural validity, (b) assuring that the measure covers the full range of the latent construct being measured, (c) eliminating redundant items, (d) maximizing sensitivity of the device used to collect responses; and (e) asking directly about change. Examples of the application of each strategy are provided. Discussion focuses on using the issues as a checklist for improving measures and the implications of sensitivity to change for prevention research with small samples.

Keywords

Sensitivity to change; Item response theory; Small sample methodology; American Indian; Alaska Native

This manuscript discusses ways prevention researchers can increase the sensitivity to change of measures in small sample studies. The central concern of prevention researchers working with small samples is poor statistical power for detecting a true preventive effect (Fok, Henry, & Allen, this issue), as power is a function of the effect size and the sample size. Detection of an effect that involves change in an outcome due to intervention or developmental effects is mediated by the ability of measures to accurately assess change, a characteristic that is variously termed “responsiveness” or “sensitivity to change” (Ebesutani, et al., 2012). However, little attention is paid in prevention research to responsiveness or sensitivity to change. A literature search of prevention journals located only a single article suggesting that sensitivity to change should be evaluated (Wästberg, Haglund, & Eklund, 2012), and no article that actually undertook such evaluation.

Although a perfectly valid instrument would be perfectly sensitive to change (Hays & Hadorn, 1992), Guyatt and colleagues suggest that, in the absence of perfect validity, responsiveness should be considered as a psychometric characteristic separate from reliability and validity. The absence of “gold standards” against which responsiveness can

be gauged has hampered research on sensitivity to change in the social and behavioral sciences (Guyatt, Walter, & Norman, 1987). Yet measures that are insensitive to change may fail to detect true effects of a prevention trial, despite the presence of other acceptable psychometric properties (Kristal, Beresford, & Lazovich, 1994), as detection of true effects is mediated by sensitivity to change. Power to detect change is reduced in proportion to sensitivity to change. That is, sensitivity to change of .5 would reduce the detectable effect size by half compared to a measure that had sensitivity to change of 1.0 (Spybrook, et al., 2009, Eq. 6.14.). Although these concerns are common challenges faced by all researchers using self-report measures to assess intervention effects, this is of particular importance in prevention studies with small samples because small sample sizes already compromise power to detect true effects.

Puhan et al. (2005) demonstrated that internal consistency reliability, as measured by Cronbach's alpha, is a poor predictor of responsiveness. Measures that distinguish among different individuals at a single point in time with high reliability may have limited utility when asked to assess change within individuals over time (Hays & Hadorn, 1992; West, Celio, Henry, & Pavuluri, 2010). Anchors or rating scales chosen for their ability to rank individuals may not be equally suited to detecting change (Ebesutani et al., 2012; Guyatt, Deyo, Charlson, Levine, & Mitchell, 1989). If respondents have difficulty differentiating among anchors close in meaning, such as "occasionally" and "sometimes," noise will be added to the measurement, possibly resulting poorer ability to detect change. In addition, a measure that is sensitive to change in one manifest indicator of a latent construct may not be sensitive to change in another indicator. For example, Wakschlag and colleagues (2014) investigated the associations between items on unidimensional scales and severity of disruptive behavior in preschoolers. They found that on some items, such as "Says 'no' when told to do something," children for whom any but the highest frequency was chosen were exhibiting normative misbehavior. However, for other items, such as "Misbehaves in ways that are dangerous," endorsement of any but the lowest frequency was associated with qualitatively atypical behavior. It is likely that change in a child who exhibited normative misbehavior would be detected by the first item, but not by the second.

Given that power is in proportion to sensitivity of change (Spybrook, et al., 2009) and that small sample prevention researchers are concerned with poor power in detecting change, it is crucial to examine the ways in which researchers increase sensitivity to change. In this article we examine five strategies for increasing the sensitivity of measures to change. We also provide examples of the application of each of these strategies in scale construction from small sample prevention studies.

Definitions and Measures of Change and Responsiveness

Kurt Lewin (1947) described change as a three-stage process, from "unfreezing", which involves overcoming inertia and dismantling the existing "mindset" as part of human surviving, to "refreezing", where a new mindset is crystallized and one's comfort level is returning to previous levels. Later models of change included Prochaska and colleagues' Transtheoretical Model of Change, which defined change as a process of five stages (Prochaska & DiClemente, 1983), as well as Bartunek and Moch's (1987) model of first,

second, and third order change focuses on cognitive schemas, which are cognitive structures that help organize and interpret a concept, in understanding individual change. Change may involve adjustments within existing schemas, modification to existing schemas, or development of capacity to change schemas. The variability in theories of change reveals complexity that may be partly responsible for the neglect of sensitivity to change in prevention research and other fields.

Various measures to defining and assessing clinically significant change may be found in the literature. Radiologists Metz (1978) along with Hanley and McNeil (1983) based a measure of change on comparison between receiver operating characteristic (ROC) curves. A 1988 consensus conference of psychiatrists defined remission in depression according to the certain degree of change on a single measure (Frank et al., 1991). Zimmerman and colleagues (2006) asked persons suffering from depression how remission should be defined, and found a wide variety of defining features. Himadi and colleagues required simultaneous improvement on multiple measures for change to be considered clinically significant (Himadi, Boice, & Barlow, 1986). The Reliable Change Index of Jacobson and Truax (1991) compared the difference between an individual's pretest and posttest scores on symptom measures to a standard error of the difference based on the reliability of the measure. Very few attempts have been made, however, to assess the potential of a measure for detecting change, which is what we mean by sensitivity to change. Proposing methods to improve sensitivity to change is the purpose of this article.

There are exceptions to this general neglect. Raudenbush and colleagues (Raudenbush & Liu, 2001; Spybrook, Raudenbush, Congdon, & Martínez, 2009) address sensitivity to change briefly as part of their discussion of hierarchical linear models for measuring change. Norman, Guyatt and colleagues have conducted extensive research on responsiveness or sensitivity to change primarily as part of a program of quality of life research (Guyatt, Walter, & Norman, 1987).

These investigations and others have produced different measures of responsiveness (Guyatt, Walter, & Norman, 1987; Liang, Fossel, & Larson, 1990; Metz, 1978; Norman, Stratford, & Regehr, 1997). In orthopedics, Liang, Fossel, and Larson (1990) compared five instruments on sensitivity to change, basing their measure of responsiveness on a comparison to a relative "gold standard" for change, namely the degree of improvement following surgery. Norman, Stratford, and Regehr (1997), however, evaluated retrospective methods based on treatment response, concluding they have limited usefulness in evaluating measures for their ability to detect change.

Guyatt, Walter, and Norman (1987) developed an indicator to evaluate the extent to which a measure is useful in assessing change over time. The Guyatt Response Index (GRI) operationalized responsiveness as the ratio of clinically-significant change to the between-subject variability in within-person change in participants. Streiner and Norman (1995) suggested using the intraclass correlation for slope as an index of responsiveness. Similarly, Spybrook and colleagues (Spybrook, et al., 2009, p. 48.) noted that the intraclass correlation for a particular polynomial of interest represents "the ability with which a researcher can

discriminate between people on their growth rate of the polynomial of interest using the least squares estimate,” or the sensitivity to change.

Where an intraclass correlation can index the degree of sensitivity to change, the significance test for the variance of linear slopes from a random effects regression model can be used to test the hypothesis that sensitivity to change is zero. The null hypothesis of this Z-test or *chi-square* test is that the variability in the estimated slopes is zero. If accepted, it suggests that the measure did not detect variability in individual change. However, it is possible to obtain a nonsignificant test because of sample characteristics rather than lack of responsiveness.

Because the variance in linear slopes in a random effects regression model is reduced to the extent that predictors such as intervention condition account for such variation, we suggest that sensitivity to change should be assessed using a mixed model that does not include variables intended to predict change, such as intervention condition. The variance component for change in such a model (usually linear slopes) provides an upper limit of intervention-related change that it will be possible to detect using the model. If a model without predictors returns a variance component for slopes that does not differ from zero, it is unlikely that the analysis will be able to detect an intervention effect on change. Thus, the power to detect change and sensitivity to change are positively related, underscoring the importance of taking measures to increase sensitivity to change in small sample prevention research.

Strategies for Increasing Sensitivity of Measures to Change

Our central focus in this section is the strategies that could be applied to increase sensitivity of measures to change in small sample prevention research, with aims to ultimately increase power to detect change. Five strategies are to: (a) Increase comprehensibility and cultural validity of the items, (b) Include items that measure the full range of the latent trait in the population, (c) Eliminate redundant and poorly-functioning items, (d) Assure that response scales function as intended, and (e) Ask directly about change. Some of these strategies will only be applicable when there is an opportunity for measurement development work prior to launching a prevention project, but others can be applied even when such development is not possible. We provide examples to demonstrate how we have applied these strategies in small sample prevention studies.

Increase comprehensibility and cultural validity

Attempting to minimize biases arising from perseveration, acquiescence, or naive hypotheses, developers of measures frequently include reverse-worded items or scale anchors (Cox, 1980; Ebesutani, et al., 2012; Schwarz, 1999). Schwarz emphasized the importance of the cognitive processes involved in survey methodology. He suggested that, in self-report questionnaires, the format, context, and the wording of questions often shape the responses obtained. The way in which respondents interpret the meaning of the questions strongly influence their responses. For example, Schwarz noted the importance of questions being developed according to Grice’s (1975) notion of a “maxim of manner,” in constructing questions. Questions should be perspicuous: Brief, avoiding ambiguity and

Author Manuscript

obscurity of meaning, and having clear order. Measures developed with college student populations are often applied to other populations or with respondents of cultural backgrounds different from those on whom they were validated (Guastello, Aruka, Doyle, & Smerz, 2008; Sharma, Beck, & Clark, 2013). Both practices may result in measures that are not comprehensible to respondents. Consider this item from an ethnic identity measure that was included in an assessment of school children: “Discrimination is not the main reason a lot of people are not as successful as they could be”, with response options, “Not at all true”, “Not very true”, “Sort of true”, and “Very true”. Attempting to avoid bias through reverse wording, the authors of this measure created an item that is wordy and confusing to the adolescent research participants completing it. Increasing comprehensibility of measures would help reduce measurement error and increase power in detecting true changes.

Author Manuscript

In prevention studies conducted with participants of non-majority cultural or linguistic backgrounds, wording that native speakers of American English take for granted may be difficult to comprehend. For example, in a small sample study Fok and colleagues (2011b) adapted the Relationship dimension of the Moos Family Environment Scale for use with Yup'ik Eskimo youth. Adaptation of the measures involves a process of cultural, linguistic, and developmental adaptation for Alaska Native adolescents. The process included successive waves of focus group discussion, item development, research team review, item revision, pilot testing, and further item revisions (Fok, Henry, Allen, Mohatt, & People Awakening Team, 2011a).

Author Manuscript

These researchers found that negatively worded items such as “Family members rarely become openly angry” had to be revised to read, “In our family we are really mad at each other a lot,” eliminating both the negative wording and the qualifier, “openly.” In another prevention study, Ebesutani and colleagues (2012) improved the psychometric properties of the Loneliness Questionnaire for administration to children and adolescence by eliminating the reverse-worded items. Additionally, they found that a three-point response scale was preferable to a five-point scale, contrary to received wisdom that having more anchors improves measures (Cox, 1980). The comprehensibility of the questions is especially important when small samples are involved, as even a single participant misinterpreting an item question could bias the measure obtained.

Author Manuscript

Attention paid to increasing the comprehensibility of items and anchors, particularly when measures are intended for use with children, will improve the ability of the measures to detect change associated with a preventive intervention, thus reducing measurement error and increasing power in detecting true intervention changes. When measures are to be used with in cultural contexts other than the contexts within which they were developed, sensitivity can be increased by attending to the cultural appropriateness of the items. Use of cultural consultants who read and comment on measures can assist in this endeavor. For example, Fok and colleagues (2011a) in their small sample prevention study developed the Multicultural Mastery Scale (MMS) for use in Alaska Native youth. The researchers found that two items, “I can get what I want by helping my friends get what they want,” and “I can get what I want by helping my family get what they want” functioned poorly. Cultural consultants suggested that these items exclusively focused on immediate self-gain, and this represented a type of individually focused motivation that was culturally undesirable to a

sample from a collective culture. The investigators eliminated these items, resulting in a scale that detected intervention-related change in a prevention trial. Thus, measures that increase comprehensibility and cultural validity may sometimes have items that are simpler in sentence structure and the meaning is obvious to research participants. Item content might also be changed based on the culture. As our intervention studies often involve small samples with culturally distinct populations that are geographically remote and difficult to access, cultural adaptation of items is a crucial step in increasing sensitivity of the measure.

Measure the full range of the latent trait

A second strategy for improving responsiveness in small samples involves assessing the extent to which the measure of a latent construct covers the full range of the construct. The range of a latent trait is the distance between the smallest and largest values of the construct in the population under study. For example, the range of symptoms of depression extends from mild dysphoria to suicidal ideation. Item response theory analysis can assist in this endeavor, by estimating the extent to which the item locations on the latent trait correspond with the distribution of the person locations. Doing this makes it possible to detect change regardless of where in the latent construct it occurs, which is important especially if the sample size is limited. Additionally, evaluating the locations of the items permits elimination of redundant items that give unequal weight to one portion of the construct. The two rug and dot plots in Figure 1 illustrate scales that do and do not cover the full range of the latent trait they are intended to assess. This simple plot, which can be created in *R* (Syntax is available from the second author upon request), can assist in evaluating measures and identifying redundant items if data on the item locations (intercepts) and individual scores on the latent trait are available.

Henry and colleagues (2008) used this method to develop a brief version of the Child Mania Rating Scale (CMRS-P; Pavuluri, Henry, Devineni, Carbray, & Birmaher, 2006) for use in small-sample clinical intervention research. One of the objectives for the development of the CMRS-P was to have high sensitivity in detecting pediatric bipolar disorder (PBD) and differentiating from attention deficit disorder hyperactivity disorder (ADHD). The full scale, which consists of 21 items, is designed to be completed by parents, unlike most mania rating scales that require clinician ratings. In the scale development study the CMRS-P was completed by parents of 150 five to 17-year-old children; 50 each with bipolar disorder, ADHD, and healthy controls. Items assess the presence and extent of DSM-IV symptoms of bipolar disorder. Example items include “Needs less sleep than usual; yet does not feel tired the next day”, and “Talks too fast and jumps from topic to topic.” Items are rated on a four-point scale ranging from “Never” to “Very Often”. The scale demonstrated excellent internal consistency ($\alpha = .96$), high accuracy in differentiating PBD from ADHD, and it covers full range of latent trait level for children with PBD.

To create a brief version, Henry et al. (2008) eliminated “redundant” items that measured the same level of the latent trait. The 10-item brief CMRS-P had internal consistency of .91, high correlation ($r = .93$) with the remaining 11 items of the full scale, and high accuracy in differentiating between PBD and ADHD. Additionally, the measure was highly sensitive to change in symptoms. West, Celio, Henry, and Pavuluri (2010) conducted a study with a

sample of 66 children with bipolar disorder, who went through a six-week treatment with medications, and were measured every week during the treatment period using the Brief CMRS-P and a clinician-rated “gold-standard” measure, the Young Mania Rating Scale (YMRS). Joint growth curve models showed that change as measured by the Brief CMRS-P correlated strongly with change as measured by the YMRS. This small sample study illustrated the importance of having non-redundant items that cover a wide range of latent trait level and that elimination of redundant items would result in a more concise measure with fewer items, assessing participants at various latent trait levels.

Eliminate redundant and poorly functioning items

Plots of item and person locations on the latent trait such as those discussed in the preceding section can assist in identifying redundant items that measure the same point on the latent continuum as other items. Although redundant items increase the internal consistency of a measure, they assign unequal weight to a single location on a scale, and, as was noted earlier, result in a measure that cannot assess change equally well across the entire range of the latent construct. This would increase measurement error and reduce power in detecting true prevention effects. Consider the latent construct of aggressiveness. A measure of aggressiveness for children might include items that range in aggressiveness from minor behaviors such as saying mean things to others through pushing and shoving, to hitting and using weapons at the highest levels. A measure such as the one depicted in the bottom panel of Figure 1, whose items are concentrated at the high (hitting and using weapons) end of the scale, would miss change in the child who desisted from minor behaviors, and would be unable to detect change that progressed from changes in minor behaviors to changes in highly aggressive but less frequent behaviors.

The sensitivity of scales to change may also be improved by eliminating items that poorly represent the construct they are designed to measure. The item discrimination index from an item response theory analysis is an indicator of the extent to which the item is able to accurately distinguish individuals at high and low levels of the latent construct. Low discrimination levels suggest items that, to return to our example of an aggression scale, would not have higher scores for more aggressive children and lower scores for less aggressive children. Inclusion of such items tends to add noise to measurement and makes change more difficult to detect.

Fok and colleagues found that eliminating poorly discriminating items improved the sensitivity to change of the Multicultural Mastery Scale (MMS; Fok et al., 2011a). This scale was developed to assess change related to a small sample cultural intervention study for rural Yup'ik youth to prevent suicide and alcohol abuse (Allen, Mohatt, Fok, Henry, & People Awakening Team, 2009; Allen, Mohatt, Fok, Henry, Burket, & People Awakening Team, 2014). The MMS was developed through a process of cultural, linguistic, and developmental adaptation of two scales for Alaska Native adolescents, the Mastery Scale (Pearlin, Menaghan, Lieberman, & Mullan, 1981), emphasizing coping through personal agency and control, and the Communal Mastery Scale (Jackson, McKenzie, & Hobfoll, 2000), that focuses on a collective approach to coping.

Samejima's (1969) graded response item response theory (IRT) model provided a slope or discrimination parameter for each item as well as location or difficulty parameters (intercepts) for each response on each item. The difficulty parameters allow us to detect redundant items and the discrimination parameters allow detection of poorly functioning items. These IRT analyses were conducted using the *ltm* package through *R* (Rizopoulos, 2006).

Two items had low discrimination indices, indicating these items were the poorest at differentiating individuals and would provide little information. These items were dropped from the final scale. The final version of the scale has been used in prevention studies in communities with samples of less than 60 (Mohatt, Fok, Henry, Allen, & People Awakening Team, 2014). Removal of redundant and poorly functioning items results in a briefer measure that discriminates participants, improve measurement precision and ultimately increase power in detecting change.

Maximize sensitivity of response options

Schwartz (1999) revealed the range of rating scales could strongly influence respondents' choice of responses. For example, Schwartz, Knäuper, Hippler, Noelle-Neumann, and Clar (1991) found that when participants were asked "How successful would you say you have been in life?", they responded significantly differently when the scale ranged from 0 to 10 or from -5 to 5, although the lowest values in both ranges represented "not at all successful" and the highest values indicated "extremely successful". This suggested that the differential interpretations of numeric value of a rating scale. A Likert-type response scale whose options are confusing or are used improperly by respondents adds noise to measurement and makes detection of change difficult, particularly when the changes involved are relatively subtle and if the sample size is small.

Thus, a fourth strategy to increase responsiveness of measures involves maximizing sensitivity of response options by evaluating the performance of response scales and optimizing them using an item response theory (IRT) analysis, such as the Samejima's (1969) Graded Response Model or the Rasch Rating Scale or Partial Credit Models (Linacre, 2002). Examining the option curves from an IRT analysis can reveal whether the coding categories in a response scale are functioning as intended. For example, in development of the MMS, Fok and colleagues (2011a) found evidence that options 2 to 4 in a five-point response scale might be providing redundant information. Figure 2 compares item characteristics curves for a five-point response scale and a three-point response scale for one of the MMS items. Note that the information provided by category 4 was almost all provided by category 3 for this item. A similar pattern was observed on all of the items, suggesting that little information was provided by having option 4, and that fewer response categories might be desirable. We recoded the data into three options by collapsing options 2–4, and reran the IRT analysis. The resulting item characteristics curve is shown in the second panel of Figure 2. The parameter estimates for the three-category calibration were very close to those from the five category calibration.

Ask directly about change

Given the importance of detecting change to prevention studies and given the difficulty in detecting change for small sample studies, it is surprising that few studies use measures that ask directly about change. Although some argue against asking directly about change (e.g., Streiner & Norman, 1995), because of difficulties in recall of previous status, such measures have been used productively in clinical research studies (e.g., Pavuluri, et al., 2004). An example is the Clinical Global Impression Scale (CGI; Guy, 1976) and its adaptation for bipolar illness (CGI-BP; Spearing, Post, Levericha, Brandt, & Nolen, 1997). The original scale was developed to provide a means for assessing the severity of symptoms of mental illness and how such symptoms change in response to psychological and pharmacological treatment. The CGI consists of three items that may be applied to any symptom or symptom profile. The first asks about severity, the second (which is not asked the first time the measure is administered) asks about improvement in symptoms, and the third asks about efficacy of pharmacological treatment.

Despite wide use in treatment studies it was years before the reliability and validity of the CGI were evaluated. When such evaluation was finally done, though validity against other measures was supported (Sato, Turnbull, Davidson, & Madakasira, 1984), the scale was criticized for unreliability and inconsistency (Beneke & Rasmus, 1992; Dahlke, Lohaus, & Gutzmann, 1992). Spearing and colleagues (1997) adapted the scale for bipolar illness, and in the process addressed several of the criticisms leveled against the scale and performed careful analysis of its psychometric properties. Where the original scale had lumped symptoms together the revised measure provided for separate ratings of mania and depression symptoms as well as an overall scale. The time domain referenced in the improvement scale was clarified and provision was made for rating symptoms that worsened. A user guide was produced that allowed for careful training of raters. Inter-rater analysis of multiple trained clinician raters and tests against a “gold standard” criterion found evidence for high sensitivity to change, suggesting the importance of asking questions of change in a measure to increase sensitivity to change.

The scale development work on the CGI suggests some rubrics for prevention scientists who wish to develop measures that ask directly about change, and can be used with small samples. First, be clear about the time frame being queried. Spearing and colleagues (1997) employed two time frames. Acute measurements were to reflect the preceding week. Prophylactic measurements were closer to the meaning of prevention, as they assessed change during the time a treatment regimen was expected to control bipolar symptoms. Second, be clear about the meaning of each response option so that respondents can identify when change has or has not occurred. Spearing and colleagues (1997), in the manual accompanying the CGI-BP, instruct raters to “integrate severity, duration, and frequency of episodes when rating change” (Spearing, et al., 1997, p. 167). Asking directly about change, however, might be suggestive to participants that change should be made, and might introduce new confounds to the survey. We suggest that readers might consider using such questions towards the end of the survey.

Discussion

Prevention research depends on the detection of change, perhaps for a treatment group for whom markers of risk or health promotion are reduced, or perhaps evidence that undesirable outcomes have increased in a control or comparison group and not in an intervention group. Measures that are weak at detecting change increase this difficulty, as decreased sensitivity of change would lead to a reduction of power in detecting true intervention change. This is particularly a concern with small sample, as power is a function of effect size and sample size. This paper has provided a review of definitions and measures of sensitivity to change, and has suggested five strategies for increasing the sensitivity of measures to change, with ultimate aim to improve measurement precision and increase power in detecting true prevention effect. These include increasing comprehensibility and cultural validity, covering the full range of the construct being measured, removing redundant items measuring the same latent trait level, optimizing the response scale, and asking directly about change. Thus, a measure that is high in sensitivity to change would consist of non-redundant items that cover a wide range of latent trait levels, and tend to be shorter in length or contain fewer items due to removal of redundant or poor functioning items. Item wordings might be simpler and item content would be tailored to the culture being assessed. Reliability of the measure might drop, as items that add little additional information are removed. The response options might be reduced compared to the original measure, as some of the response categories might provide little additional information. It should also contain items that ask directly about change, and will contain items that are obvious in meaning and easily interpreted by participants.

Not all of the strategies used in this paper will be useful in any given study or with any particular measure. They are particularly important considerations for the main outcomes and key mediators in a trial where detecting change is of utmost importance. Each strategy will only be useful if the applicable pre-existing characteristics of the measure limits sensitivity to change. For example, reducing the number of anchors is not likely to increase sensitivity unless the existing anchors do not provide unique information, as was the case in the example given. The same is true of eliminating redundant items. The issues raised in this article should be considered when selecting or adapting measures for a prevention trial with a small sample.

The issues raised in this paper may be used, roughly in the order presented, as a checklist of considerations in selecting measures for a prevention trial. Of first importance is that a measure be understood by those who complete it. Numerous measures administered in studies with children were developed and validated on college students, with unfortunate and sometimes humorous results. For example, the Metropolitan Area Child Study employed a measure of aggressive strategies that presented a respondent with a situation such as a conflict over a possession. The instructions asked the respondent to rank four responses (hit, yell, talk, do nothing) in order of the likelihood that the respondent would employ them. Originally developed on college students, this measure had been adapted for first through sixth grade children by modifying the situations and simplifying language. When some children were interviewed after completing an assessment packet containing this measure, it became apparent that children did not understand likelihood in the same sense

employed in the measure. A typical response was, “First I would say something, then I would yell at him, and then I would hit him!”

A measure that can be understood by participants and covers the full range of the construct to be measured can be improved further by eliminating items that have the same degree of difficulty in IRT terms, i.e., that measure the same level of the latent construct. Reliance on internal consistency as an index of reliability has encouraged the development of measures with many redundant items, often leading respondents to complain that the investigator seems to be “asking the same question over and over.” When measuring change, a measure made up of many redundant items will be more susceptible to the effects of repeated testing on the internal validity of study (Campbell, Stanley, & Gage, 1963), as participants may be likely to remember previous responses to items that seem to be repeated.

The last two issues discussed in this paper involve improvements to the device used to collect responses: Optimizing an existing response scale by assuring that each option contributes unique information, or asking directly about change. We believe that use of these strategies can improve the sensitivity of measures to change by reducing noise and/or tightening the link between the change in the phenomenon of interest and the response.

To summarize, we offer these suggestions in the belief that attention to the issue of sensitivity to change is neglected in prevention science as it is in psychosocial research generally. These strategies would be particularly useful in small sample prevention studies in detecting change. Whether or not the validity of a measure is also an index of its sensitivity to change (Hays & Hadorn, 1992), it is certainly the case that reliability and responsiveness are distinct aspects of measurement (Guyatt, et al., 1987). The potential cost to prevention science of employing measures with low sensitivity to change combined with the potential for improving sensitivity to change argue for greater attention to this aspect of measurement.

References

- Allen J, Mohatt G, Fok CCT, Henry D. People Awakening Team. Suicide prevention as a community development process: Understanding circumpolar youth suicide prevention through community level outcomes. *International Journal of Circumpolar Health*. 2009; 68:274–291. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875412/>. [PubMed: 19705659]
- Allen J, Mohatt GV, Fok CCT, Henry D, Burket R. People Awakening Team. A protective factor model for alcohol abuse and suicide prevention among Alaska Native youth. *American Journal of Community Psychology*. 2014; 54:125–139.10.1007/s10464-014-9661-3 [PubMed: 24952249]
- Bartunek JM, Moch MK. First-Order, Second-Order, and Third-Order Change and Organization Development Interventions: A Cognitive Approach. *The Journal of Applied Behavioral Science*. 1987; 23:483–500.10.1177/002188638702300404
- Beneke M, Rasmus W. ‘Clinical global impressions’ (ECDEU): some critical comments. *Pharmacopsychiatrie*. 1992; 25:171–176.
- Campbell, DT.; Stanley, JC.; Gage, NL. *Experimental and quasi-experimental designs for research*. Boston, MA, US: Houghton, Mifflin and Company; 1963.
- Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. 2. New York, NY: Lawrence Erlbaum Associates; 1988.
- Cox EP. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*. 1980; 17:407–422.10.2307/3150495

- Dahlke F, Lohaus A, Gutzmann H. Reliability and clinical concepts underlying global judgments in dementia: Implications for clinical research. *Psychopharmacology Bulletin*. 1992; 28:425–432. [PubMed: 1296220]
- Ebesutani C, Drescher CF, Reise SP, Heiden L, Hight TL, Damond JD, Young J. The Loneliness Questionnaire–Short Version: An Evaluation of Reverse-Worded and Non-Reverse-Worded Items Via Item Response Theory. *Journal of Personality Assessment*. 2012; 94:427–437.10.1080/00223891.2012.662188 [PubMed: 22404209]
- Fok CCT, Allen J, Henry D, Mohatt G. People Awakening Team. Multicultural Mastery Scale for Youth: Multidimensional assessment of culturally mediated coping strategies. *Psychological Assessment*, Advance online publication. 2011a Sep 19.10.1037/a0025505
- Fok CCT, Allen JA, Henry D. The People Awakening Team. The Brief Family Relationship Scale: A Brief Measure of the Relationship Dimension in Family Functioning. 2011b Nov 14. Advance online publication. 10.1177/1073191111425856
- Fok CCT, Henry D, Allen JA. Maybe Small is too Small a Term: Introduction to Advancing Small Sample Prevention Science. this issue.
- Frank E, Prien RF, Jarrett RB, Keller MB, Kupfer DJ, Lavori PW, Rush AJ, Weissman MM. Conceptualization and rationale for consensus definitions of terms in major depressive disorder. *Archives of General Psychiatry*. 1991; 48:851–855. [PubMed: 1929776]
- Grice, H Paul. Logic and Conversation. In: Cole, P.; Morgan, JL., editors. *Syntax and Semantics*. Vol. 3. New York: Academic Press; 1975. p. 41–58.
- Guy, W. Clinical Global Impressions. In: Guy, W., editor. *ECDEU Assessment Manual for Psychopharmacology*, revised. Rockville, MD: National Institute of Mental Health; 1976. p. 218–222.
- Guastello SJ, Aruka Y, Doyle M, Smerz KE. Cross-cultural generalizability of a cusp catastrophe model for binge drinking among college students. *Nonlinear Dynamics Psychology, and Life Sciences*. 2008; 12:397–407.
- Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*. 1987; 40:171–178. [PubMed: 3818871]
- Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. *Journal of clinical epidemiology*. 1989; 42(5):403–408. [PubMed: 2659745]
- Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology*. 1983; 148:839–843. [PubMed: 6878708]
- Hays RD, Hadorn D. Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research*. 1992; 1:73–75. [PubMed: 1301117]
- Henry D, Keys CB, Schaumann L. Value-based job analysis: An approach to human resource management in rehabilitation agencies serving people with developmental disabilities. *Journal of Rehabilitation Administration*. 2001; 25:1–17.
- Henry D, Pavuluri MN, Youngstrom E, Birmaher B. Accuracy of Brief and Full Forms of the Child Mania Rating Scale. *Journal of Clinical Psychology*. 2008; 64:1–14.10.1002/jclp.20464 [PubMed: 18161043]
- Himadi WG, Boice R, Barlow DH. Assessment of agoraphobia—II: Measurement of clinical change. *Behaviour Research and Therapy*. 1986; 24(3):321–332. [PubMed: 3729903]
- Huesmann LR, Guerra NG. Children's normative beliefs about aggression and aggressive behavior. *Journal of Personality and Social Psychology*. 1997; 72:408–419. [PubMed: 9107008]
- Jackson, T.; McKenzie, J.; Hobfoll, SE. Communal aspects of self-regulation. In: Boekaerts, M.; Pintrich, PR.; Zeider, M., editors. *Handbook of self-regulation*. San Diego, CA: Academic Press; 2000.
- Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of consulting and clinical psychology*. 1991; 59(1):12. [PubMed: 2002127]
- Kristal AR, Beresford SA, Lazovich D. Assessing change in diet-intervention research. *The American journal of clinical nutrition*. 1994; 59(1):185S–189S. [PubMed: 8279421]
- Lewin K. Frontiers in group dynamics. *Human Relations*. 1947; 1:143–153.

- Liang MH, Fossel AH, Larson MG. Comparisons of five health status Instruments for orthopaedic evaluation. *Medical Care*. 1990; 28:632–642. [PubMed: 2366602]
- Linacre JM. Optimizing rating scale category effectiveness. *Journal of Applied Measurement*. 2002; 3:85–106. [PubMed: 11997586]
- Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*. 1978; 8:283–298. [PubMed: 112681]
- Mohatt GV, Fok CCT, Henry D, Allen J. People Awakening Team. Feasibility of a community intervention for the prevention of suicide and alcohol abuse with Yup'ik Alaska Native youth: The Elluam Tungiinun and Yupiucimta Asvairtuumallerkaa studies. *American Journal of Community Psychology*. 2014; 54:153–169.10.1007/s10464-014-9646-2 [PubMed: 24952248]
- Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach alpha. *Journal of Clinical Epidemiology*. 1997; 50:869–879. [PubMed: 9291871]
- Pavuluri MN, Henry D, Naylor M, Carbray J, Sampson G, Janicak PG. A Pharmacotherapy Algorithm for Stabilization and Maintenance of Pediatric Bipolar Disorder. *Journal of American Academy of Child and Adolescent Psychiatry*. 2004:859–867.
- Pavuluri MN, Henry DB, Devineni B, Carbray JA, Birmaher B. Child Mania Rating Scale: Development, reliability, and validity. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2006; 45:550–560.10.1097/01.chi.0000205700.40700.50 [PubMed: 16601399]
- Pearlin LI, Menaghan EG, Lieberman MA, Mullan JT. The stress process. *Journal of Health & Social Behavior*. 1981; 22:337–356. [PubMed: 7320473]
- Puhan MA, Bryant D, Guyatt GH, Heels-Ansdell D, Schünemann HJ. Internal consistency reliability is a poor predictor of responsiveness. *Health and Quality of Life Outcomes*. 2005; 3(33):1–8.10.1186/1477-7525-3-33 [PubMed: 15634354]
- Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: Toward an integrative model of change. *Journal of Consulting and Clinical Psychology*. 1983; 51:390–395. [PubMed: 6863699]
- Rasch G. An item analysis which takes individual differences into account. *The British Journal of Mathematical and Statistical Psychology*. 1966; 19:49–57. [PubMed: 5939145]
- Raudenbush SW, Liu X. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*. 2001; 6:387–401.10.1037/1082-989X.6.4.387 [PubMed: 11778679]
- Rizopoulos D. ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*. 2006; 17:1–25.
- Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969; 34:100–114.
- Sato TL, Turnbull CD, Davidson JRT, Madakasira S. Depressive illness and placebo response. *International Journal of Psychiatry Medicine*. 1984; 14:171–179.
- Schwarz N. Self-reports: How the questions shape the answers. *American Psychologist*. 1999; 54:93–105.
- Schwarz N, Knäuper B, Hippler HJ, Noelle-Neumann E, Clar F. Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*. 1991; 55:570–582.
- Sharma E, Beck KH, Clar PI. Social context of smoking hookah among college students: scale development and validation. *Journal of American College Health*. 2013; 61:204–211.10.1080/07448481.2013.787621 [PubMed: 23663124]
- Spearing MK, Post RM, Levericha GS, Brandt D, Nolen W. Modification of the Clinical Global Impressions (CGI) scale for use in bipolar illness (BP): the CGI-BP. *Psychiatric Research*. 1997; 73:159–171.
- Streiner, DL.; Norman, GR. *Health measurement scales: A practical guide to their development and use*. 3. Oxford: Oxford University Press; 1995.
- Spybrook, J.; Raudenbush, SW.; Congdon, R.; Martínez, A. *Optimal design for longitudinal and multilevel research: Documentation for the “Optimal Design” software*. New York, NY: William T. Grant Foundation; 2009.

- Wakschlag LS, Briggs-Gowan MJ, Choi SW, Nichols SR, Kestler J, Burns JL, Henry D. Advancing a Multidimensional, Developmental Spectrum Approach to Preschool Disruptive Behavior. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2014; 53(1):82–96. [PubMed: 24342388]
- Wakschlag LS, Henry DB, Tolan PH, Carter AS, Burns JL, Briggs-Gowan MJ. Putting theory to the test: modeling a multidimensional, developmentally-based approach to preschool disruptive behavior. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2012; 51(6):593–604. [PubMed: 22632619]
- Wästberg BA, Haglund L, Eklund M. The work environment impact scale–self-rating (WEIS-SR) evaluated in primary health care in Sweden. *Work: A Journal of Prevention, Assessment and Rehabilitation*. 2012; 42(3):447–457.
- West AE, Celio CI, Henry DB, Pavuluri MN. Child Mania Rating Scale-Parent Version: A valid measure of symptom change due to pharmacotherapy. *Journal of Affective Disorders*. 2011; 128:112–119. [PubMed: 20858565]
- Zimmerman M, McGlinchey JB, Posternak MA, Friedman M, Attiullah N, Boerescu D. How should remission from depression be defined? The depressed patient’s perspective. *American Journal of Psychiatry*. 2006; 163(1):148–150. [PubMed: 16390903]

Figure 1.

Illustrations of scales whose items do and do not cover the full range of a latent trait as it is found in the population. Circles represent items and the rug plot represents persons. The bottom scale would be sensitive to change only in the highest ranges of the latent trait, whereas the top scale could detect change at any level of the latent trait.

A

B

Figure 2. Comparison of trace lines for the five- (A) and three-category (B) calibrations for an item in the Multicultural Mastery Scale (Fok et al., 2011a).