Description and Psychometric Evaluation of the Youth Counseling Impact Scale

Manuel Riemer and
Department of Psychology, Wilfrid Laurier University

Marcia A. Kearns
Department of Psychological Sciences, University of Missouri - Columbia

Abstract

The current paper describes and evaluates the Youth Counseling Impact Scale (YCIS), a recently developed therapeutic process measure that assesses youths' perceptions of the impact individual mental health counseling sessions have on their thoughts, feelings, and behaviors. This measure is intended for frequent use in the mental health treatment of youths aged 11–18. It provides a general Impact score as well as two subscale scores: Insight and Change. Five hundred youths receiving mental health services participated in this investigation. Classical Test Theory, Item Response Theory, Confirmatory Factor Analysis, and analyses of the relationship of the YCIS with other scales were used to evaluate the research questions. The results suggest that overall the YCIS is a well-functioning scale with good psychometric properties. The proposed model for one primary general factor of impact and two secondary factors (Insight and Change) fit the data well. Specific weaknesses of the scale are discussed and possible improvements are suggested.

Keywords

counseling impact; session impact; therapeutic processes; youths; measurement

Description and Psychometric Evaluation of the Youth Counseling Impact Scale

Background

Currently, there is little research on the processes that underlie change in youth mental health treatments. The lack of inquiry in this area limits our understanding of why or how youth treatments work (Kazdin & Nock, 2003). This paper describes and evaluates the psychometric properties of the Youth Counseling Impact Scale (YCIS), a recently developed therapeutic process measure that may provide insight into the perceived impact of individual counseling sessions on youths' thoughts, feelings, and behaviors.

The YCIS was developed and tested as a component of the Peabody Treatment Progress Battery, a comprehensive battery of short, reliable, and efficient instruments designed for frequent use with youths receiving mental health services (Bickman, Riemer, Lambert, et al, 2007). This battery was created to offer feedback on therapy processes and outcomes to counselors, and includes both English and Spanish versions. The YCIS, specifically, is a therapeutic impact measure, developed for use in individual treatment sessions with youths aged 11–18. Generally, measures of impact, "Are concerned with clients' internal reactions to sessions, which, logically, must intervene between in-session events and the long-term effects of treatment" (Stiles et al., 1994, p.175). These measures provide information about clients' perceptions of the effects of therapy on a session by session basis and, thus, provide a useful way of measuring the progression of treatment.

Other Session Impact Measures

The most commonly used session impact measures, the Session Evaluation Questionnaire (SEQ) (Stiles, 1980) and the Session Impact Scale (SIS) (Elliot & Wexler, 1994), were both designed to be used with adults. The SEQ consists of several bipolar adjective pairs that describe perceptions of session impact on two dimensions: depth and smoothness (Stiles, 1980). The depth dimension, "distinguishe[s] sessions that [are] deep, valuable, full, special, and good from those that [are] shallow, worthless, empty, ordinary, and bad" (Stiles, 1980, p. 181). The smoothness dimension "distinguishe[s] sessions that [are] smooth, easy, pleasant, and safe from those that [are] rough, difficult, unpleasant, and dangerous" (Stiles, 1980, p. 181). The SIS assesses perceptions of session impact on the dimensions of hindering versus helpful impacts, with helpful impacts having two sub-dimensions: task impacts and relationship impacts (Elliot & Wexler, 1994). The SEQ focuses on general emotional reactions to sessions while the SIS targets the specific content of therapy sessions (Elliot & Wexler, 1994).

Session Impact and Adults—Studies with adults have revealed that ratings of session impact relate to a variety of therapeutic constructs and patient and therapist variables. For example, ratings of session impact have been found to vary as a function of therapy type and treatment stage (Reynolds et al. 1996). Ratings of session impact have also been found to relate to personal characteristics such as racial group membership of clients and therapists (Gregory & Leslie, 1996), client introversion (Kivlighan & Angelone, 2001; Nocita & Stiles, 1986), attachment style (Mohr, Gelso, & Hill, 2005), and symptom severity (Odell & Quinn, 1998). Ratings of session impact have also been found to be associated with clients' perceptions of counseling helpfulness (Barak & Bloch, 2006) and working alliance (Fitzpatrick, Iwakabe, & Stalikas, 2005). Finally, ratings of session impact have been found to be associated with client involvement and progress in treatment (Eugster & Wampold, 1996), client dropout rates, (Samstag, Batchelder, Muran, Safran, & Winston, 1998) and client return rates (Tryon, 1990).

Session Impact and Youths—The few studies that have examined ratings of session impact in youths have produced similar findings to those done with adults. For example, a study of an adolescent group intervention targeting intimacy and identity found that ratings on a measure of session impact reflected differences in the therapeutic processes between therapy groups (Bussell, 2001). An exploratory study of what adolescent males found to be helpful in counseling found that they showed similar patterns of session ratings to adults and rated sessions where they obtained insight as especially helpful (Dunne, Thompson, & Leitch, 2002). A study, comparing differences in youths' ratings of session impact and therapeutic alliance in telephone versus online counseling, found that ratings of session impact predicted client outcomes and varied as a function of the type of counseling received (King, Bambling, Reid, & Thomas, 2006). Lastly, a study examining the relationship between session impact and the

development of working alliance in youths undergoing counseling found that session impact was related to the formation of the working alliance (Ji, 2002).

The Need for a New Measure—Measures of session impact designed specifically for youths, however, are lacking. A review of the current research literature revealed only one impact measure developed for use with adolescents: the Session Evaluation Form (SEF). The SEF is an unpublished measure that was adapted from the SIS and developed specifically for group therapy sessions (see Bussell, 2001). Group therapy sessions, however, have been found to differ in therapeutic processes from individual counseling sessions (Holmes & Kivlighan, 2000). Additionally, the SEF (like most session impact scales) only assesses the internal reaction immediately after a session and not the behavioral changes that occur in the weeks following a session. These changes, however, may provide relevant information about whether youths are using what they learn in therapy to change their behaviors. In contrast, the YCIS was developed for use in individual counseling sessions and provides information about both the internal reaction following a session and the behavioral changes that occur in the weeks following a session.

Description of the YCIS

In sum, the YCIS measures youths' perceptions of helpful counseling impact, the perceived immediate effect of a counseling session on clients' understanding of their feelings, relationships, and problems (insight), and the positive behavioral and emotional changes that occur in the weeks immediately following a counseling session (change). The YCIS contains ten items worded appropriately for youths. The first five, derived from the task impact dimension described by Elliot and Wexler (1994), assess insight immediately after a session. These are: *I learned something important about myself; I now understand better something about somebody else (like my parents, friends, or my brother or sister); I now understand my feelings better; I now understand better what my problems may be; I now have a better idea about how I can deal with my problems*. The second five measure the emotional and behavioral changes that occurred in the two weeks prior to the current session. They are: *I tried things my counselor suggested; I felt better about myself than before; I used things that I learned in counseling; I improved my behavior in my home; I improved my behavior in school, at work, or other places like these outside of my home*. Elliot and Wexler (1994) also described a hindering dimension, which they define as "clients' negative experiences" in therapy (p. 166–167), which we opted not to include. While we recognize that there may be a need to assess such negative experiences in therapy, we had several reasons to not include this dimension in this scale: 1) We were primarily interested in the positive impact of therapy (or lack thereof); 2) a separate therapeutic alliance scale captures negative experiences of clients in treatment quite well; 3) we wanted to keep the scale as short as possible; and 4) the psychometric evidence presented by the authors did not convince us that this dimension could easily be integrated into an overall scale of session impact.

Each item is rated on a 5-point scale (*1= Not At All, 2= Only a little, 3= Somewhat, 4 =Quite a bit, 5= Totally*) and the recommended frequency of administration is every two weeks. The purpose of this study is to evaluate the psychometric properties of the Youth Counseling Impact Scale. Based on our theoretical conceptualization of counseling impact we expect a hierarchical G-Factor model with one primary general factor (counseling impact) and two secondary factors (insight and change).

Method

Sample

The sample was drawn from a large national psychometric study, conducted in the summer of 2005, with youths aged 11–18, their caregivers, and their counselors. The participants were recruited from 28 regional offices in six different states (FL, IL, IN, MI, TX, VA) comprising part of a large national provider for home-based mental health services. The final sample used for this analysis consisted of 500 youths. The gender and racial/ethnic breakdown of the sample is as follows: 41% female, 64% White, 32% African American, and 18% of Hispanic descent. The average age was 14. Almost all (96%) reported English as their primary language. Many of the youths were court referred to these services, with over half (58%) having been arrested at least once. Of those arrested, 64% reported having been arrested during the last 12 months. About two thirds (66%) had either currently or previously been diagnosed with a mental health disorder. The large majority (90%) had been in treatment for at least a month and one third had been in treatment for more than six months. Half of the youths in the sample lived with their biological parents, a third (32%) with foster parents, and the remainder had different living arrangements.

Procedure

The youths completed a background questionnaire, a series of outcome measures, and a set of measures designed to assess therapeutic variables believed to influence treatment outcomes (e.g., therapeutic alliance, treatment motivation, expectancies). These measures were divided into two booklets that were administered by the youth's counselor at the end of two sessions. The background and the outcome measures were completed in the first session and all other measures were completed in the next session. The YCIS was part of the second booklet. The de-identified completed questionnaires were returned to our research center (the Center for Evaluation and Program Improvement) and entered following a rigorous data processing protocol (see Bickman et al., 2007). For the current study, we included youths who had at least 85% of the YCIS completed. This study was approved by the Institutional Review Board of Vanderbilt University.

Analyses

For the evaluation of the psychometric properties of this scale, we used currently available models for psychometric analyses, namely Classical Test Theory (CTT), Item Response Theory (more specifically: Rasch measurement), and Confirmatory Factor Analysis (CFA). Each of the models produced indicators for the psychometric quality for the overall scale as well as for each item. These figures made it possible to identify stronger and weaker items in a given test. An analysis of its relationship to other variables was also conducted. The plan for additional studies providing validity evidence will be described in the conclusion section.

Classical Test Theory—The goal of using CTT was to provide familiar indicators of the psychometric characteristics of the YCIS. We will describe the total scale as well as the subscales with summary statistics and an indicator of the internal reliability of its test scores (i.e., Cronbach's coefficient alpha). We will also investigate the characteristics of each item by investigating its distributional characteristics and its relationship to the overall scale and subscale. The strength of this approach is its ease of use and most readers' familiarity with it. The statistics, however, are sample dependent and have arithmetic operations that require variables at the interval scale level. Interval level scaling, however, cannot be assumed for rating scale items without further investigation and manipulation, such as using a Rasch measurement model (Wright & Mok, 2004).

Rasch Measurement—Item response theory has the advantage of providing more detailed item level information that is less dependent on the specific sample than classical test theory. In this approach, measurement models that estimate both item level and person level parameters are used to create linear interval-level scales (Embretson, 1996). Several item response models have been developed over the years following the lead of Lord and Novick (1968) as well as Rasch (1960). In this paper, we applied the Rasch measurement approach. Of the family of Rasch measurement models, the appropriate one for creating measurement scores from the YCIS raw scores is the rating scale model (Andrich, 1978; Wright & Masters, 1982; Wright & Mok, 2004). Besides the person ability (Bn) and the item difficulty score (Di), the rating scale model also includes a difficulty estimate (Fx) for the item thresholds. The item threshold is the person measure score expressed in logit units for which a person with that score is equally likely to endorse one answer category over another (e.g., *Only A Little* over *Not At All*). The general Rasch rating scale model is described by the following equation:

$$\ln\left(\frac{P_{nix}}{P_{nix-1}}\right) = B_n - D_i - F_x$$

(1)

In this equation, the natural logarithms of the odds ratio of the probability P of person n choosing category x of item i over selecting the previous category (e.g., the probability of selecting *Only a Little* over *Not at All*) is modeled as the difference of the person ability or trait (Bn; e.g., the level of perceived impact), the item difficulty (Di; i.e., how likely it is that a respondent would endorse this item at a high level), and each threshold estimate (Fx). This additive functionality plus the fact that the estimated parameters are sample independent are great advantages of the Rasch measurement approach.

In conducting the analysis and reporting the results, we followed the guidelines established for the Journal of Applied Measurement (Smith, Linacre, & Smith, 2003), one of the main scholarly outlets for Rasch measurement work. For creating measurement scores and evaluating the fit of the data to the Rasch model, we used Winsteps (Version 3.61.1), a commonly used Rasch measurement software developed by Linacre and Wright (2000). Winsteps uses Joint Maximum Likelihood Estimation which is explained in Linacre (2004). The results of this analysis will give us some important insight into the psychometric quality of the YCIS as a scale.

Confirmatory Factor Analysis—We constructed the YCIS to function with a hierarchical factor structure, so that a total impact scale score could be used as well as two secondary factors (insight and change) that both significantly contribute to the higher order factor. In order to provide evidence for this model we used Confirmatory Factor Analysis (Bollen, 1989) to test whether the model has a good fit with the current data. The model we tested is depicted in Figure 1. Items 1–5 are modeled to load onto the insight factor and items 6–10 on the change factor. We constrained the loadings of the secondary factors onto the general factor to be equal and the variance of the general factor to be 1 in order to prevent an over-identified model. To estimate the models and their fit with the data we used the SAS procedure PROC CALIS.

Since confirmatory analysis can only demonstrate that the current model fits the current data reasonably well, but not whether it is the model that would best explain the variance-covariance structure in the data, we conducted two additional tests. First, we tested the fit of another plausible model in which all items load directly on just one general impact factor. If this simpler model fits the data equally well, then the more complex hierarchical G-factor model may not be justified. These two models can be directly compared using the Chi Square difference test since constraining both loadings (a) of the secondary factor onto the general factor to be equal

to one is mathematically the same as testing a one-factor model. Thus, the two models are nested within each other.

Second, we used the results of the dimensionality investigation that is part of the Rasch measurement construction to provide further evidence for the proposed factorial structure. For the dimensionality investigation one first creates measurement scores from observational data and then uses exploratory principal component factor analysis of the standardized residuals to test whether there are plausible additional factors beyond the one that is represented by the measure scores (Linacre, 1998; Smith, 2004).

Relationship to other scales—The perceived impact of the counseling session on youths is just one aspect of the therapy or counseling process. Other common factors such as therapeutic alliance and treatment motivation are also important parts of the process (Karver, Handelsman, Fields, & Bickman, 2005). From a theoretical standpoint one would expect that counseling impact is related to these common factors but also distinct. As discussed in the introduction, ratings of session impact have been found to be positively associated with clients' perceptions of counseling helpfulness (Barak & Bloch, 2006). Perceived helpfulness is a concept closely related to overall client satisfaction. Thus, we expect that scores from a service satisfaction scale will be closely related to (in a positive direction) but still distinct from the YCIS scores. Similarly, session impact has been, theoretically and empirically, assumed to be related to working alliance (Elliot & Wexler, 1994; Fitzpatrick, Iwakabe, & Stalikas, 2005). It seems intuitive that counselors who have a stronger alliance with their clients are more likely to have a positive impact on them. For this reason we hypothesize that the scores from the YCIS will be strongly and positively related to ratings of the working alliance assessed at the end of a session. However, these scores should also be empirically distinct from each other since session impact is not the same as the working alliance. Finally, ratings of session impact have been found to be associated with client involvement and progress in treatment (Eugster & Wampold, 1996), client dropout rates, (Samstag, Batchelder, Muran, Safran, & Winston, 1998) and client return rates (Tryon, 1990) as mentioned earlier. All of these concepts are related to treatment motivation. Consequently, the YCIS scores should be related to (in a positive direction) but distinct from the scores of a measure of treatment motivation. We will provide some preliminary investigations into whether there is empirical evidence for the above hypothesized relationships providing further information regarding the validity of using the scale scores for the intended purposes. We used the Youth Therapeutic Alliance Quality Scale (Y-TAQS), the youth version of the Motivation for Youth Treatment Scale (Y-MYTS), and a brief youth version of a service satisfaction scale (Y-SSS). All three scales are part of the Peabody Treatment Progress Battery and their psychometric properties have been described in the corresponding manual (Bickman, et al., 2007).

We investigated the relationships between these scales and how distinct they are by calculating the correlations between these constructs. If one of these correlations was greater than what is considered large according to Cohen's (1992) standards, that is, $|r| > .50$ (i.e., 25% of the variance is shared), we further investigated the discrimination by studying the correlation matrix of all items in the two scales in question. If the two scales represent two distinct constructs, a pattern should be clearly discernable in which the items within a scale correlate higher than items between scales. It needs to be noted that a more elaborate discriminate validity study would be preferable but is beyond the scope of this paper. Future publications are planned for this purpose.

Results

Classical Test Theory

Total Scale—Each of the 10 YCIS items was rated on a 5-point scale. To build the raw YCIS total score a simple average of each youth's ratings across all items is calculated. Thus, the total mean raw score ranged from 1 to 5. The distribution of YCIS total scores in the current sample has a mean of 3.59 and a standard deviation of 1.00. Overall, the scores are normally distributed. The distribution is slightly skewed to the left (−.48) and has only a small negative kurtosis (−.38). The median is close to the mean with a value of 3.67. Coefficient alpha for the total score is high at .92 with generally high item total correlations ranging from .59 (item 10) to .80 (item 3). The means for the individual items range from 3.26 (item 10) to 3.81 (item 5). Their standard deviations are between 1.22 and 1.56. All items are only slightly skewed and have small kurtoses.

Insight Subscale—Similar to the raw total YCIS score, the raw score for the Insight subscale is calculated by taking the mean of the ratings for items 1 through 5. The distribution of the Insight scores is also approximately normal. The mean (3.66) is slightly higher than the YCIS total score as is the median (3.80). The distribution is slightly skewed to the left (−.65) and has a small negative kurtosis (−.27). The Cronbach's alpha for this subscale is .91 with item-total correlations ranging from .73 to .83.

Change Subscale—The raw score for the Change subscale is represented by the mean of items 6 through 10 and is approximately normal as well. The mean (3.48) is slightly lower than the YCIS total score while the median is almost the same (3.60). The distribution is also slightly skewed to the left (−.40) and has a small negative kurtosis (−.51). The coefficient alpha for this subscale is .86 with item-total correlations ranging from .59 to .79.

Rash Measurement Model

The item characteristic curve shown in Figure 2 below shows the category probabilities for each of the answer choices. For an ideal scale, each category (1 through 5) has the highest probability of being endorsed at some point along the person score continuum (Bond & Fox, 2001). This is accomplished for the first category at the left end of the X-Axis, for the third and fourth in the middle, and for the fifth on the right end. The second answer category (O*nly A Little*) does not reach that point. The item thresholds are at −.45 (1 to 2), −1.29 (2 to 3), .38 (3 to 4) and 1.36 (4 to5) logit units, which is where the probability curves cross. That means, for example, a person with an ability score of −.45 logit units is just as likely to endorse category 1 (*Not At All*) as they would endorse category 2 (*Only A Little*). Ideally these thresholds increase in a monotone way. Again this is not the case for the threshold of categories two and three (−1.29) which is lower than the one from one to two (−.45). These results suggest that respondents have difficulties differentiating between *Only A Little* and *Somewhat*. One could consider revising the scale and only using four answer categories. However, because the YCIS is part of a measurement battery where most of the instruments use the same five-point ratings, we opted to keep the five answer categories for the YCIS to avoid respondents from having to switch back and forth between four and five answer categories.

One of the first steps in analyzing characteristics of a scale in the Rasch measurement approach is to examine Table 1 and 2, which provide a summary of the key model fit statistics. The first statistics to consider are the person and item reliability scores. The person reliability of .83 suggests that the scale scores discriminate reasonably well between persons. The person reliability is approximately equivalent to coefficient alpha so that values above .80 are considered satisfactory (Clark & Watson, 1995). The item reliability of .93 indicates that the items of the YCIS create a well defined variable.

The next step is to evaluate the location of the persons relative to the items. The default is to set the item mean to 0, which we did. The mean person score is .47 logit units and .89 if the extreme observations are included[1]. This implies that it is easier for a youth to endorse high scores compared to a perfectly balanced item-to-person situation in which the mean person score would be lined up with the mean item score at 0. High scores in this context refer to an indication of a strong positive impact of counseling as reported by the youth. The person map of items in Figure 3 shows that overall the items and person are well lined up with the slight shift upwards described above. This graph also illustrates that there are no items that represent the extreme low or high scoring youths. Thus, the YCIS is most precise in differentiating cases in the middle of the score range and will be less sensitive to changes at the extremes. This means the difference between the extreme low true scores of 1 and 1.1, for example, is less precisely measured compared to the difference between 2.5 and 2.6. This is important to keep in mind when interpreting the scores of the YCIS.

Checking the residual matrix of the expected scores relative to the observed scores is also common practice. There are two types of fit statistics that provide summary information about the degree of deviation and misfit: outfit and infit. Outfit is based on the conventional sum of squared standardized residuals while infit is an information-weighted sum. The outfit statistic is more sensitive to extreme observations, while the weighting lessens the influence of those for the infit statistics. It is common to report the outfit and infit statistics as means square values (MNSQ), as well as standardized values (ZSTD)--a type of t-statistic. If the data fit the model perfectly, the t-statistics should have a mean of 0 and a standard deviation of 1. T values greater than +2 or less than −2 are generally interpreted as having less compatibility with the model than expected (Bond & Fox, 2001). That means there are several people who, based on their general scoring behavior, should have endorsed this item in a certain expected way, but did not. There are no hard-and-fast rules for interpreting the MNSQ. However, Wright and Linacre (1994) suggest as a reasonable item mean square range for rating scales .6 to 1.4 (the potential range is 0 to infinity) with the expected value being 1.

Using these general guidelines it is clear that both the general item and person fit are very good for the YCIS (see Table 1 and 2). The person infit MNSQ is 1.07 (SD = .89) and the ZSTD is −.30 (2.00) while the respective outfit values are almost the same at 1.06 (.87) and −.30 (2.00). The outfit MNSQ of 1.06 indicates that there is only 6% more noise than modeled. The standard deviations are a little bit higher than the expected value which is due to some unusual observations. The item fit indices are as follows: infit MNSQ = 1.01 (.29), ZSTD = −.2 (4.0), and the corresponding outfit values are 1.06 (.30) and .5 (3.7). Again, the standard deviations are higher than ideal.

This elevated standard deviation is partly due to the less than ideal fit of item 10 (Outfit MNSQ = 1.75). This item was phrased: "*In the last two weeks I have improved my behavior in school*." Since the psychometric study was conducted during the summer when most youths were out of school, many of them probably had difficulty answering this item. This explains the unusual pattern and variance for that item and the less than ideal fit with the overall scale. This result suggests that this item should either be dropped or reworded. In the most recent version of the YCIS (that is available as part of the Peabody Treatment Progress Battery), the item has been reworded to include others places besides school and now reads: "I improved my behavior in school, at work, or other places like these *outside of my home*."

Extreme observations such as a five on all 10 items or a 1 on all items are typically excluded from the analysis because they correspond to indefinite measures and can take on any value outside the measurement range of the test. However, procedures have been developed to include extreme scores in the calculation of certain summary statistics (Linacre, 2004).

Another aspect of the fit of the data to the Rasch model is unidimensionality. While it is clear that empirical data are always manifestations of more than one latent dimension, in the Rasch measurement framework it is important to demonstrate that there is only one primary latent variable that is represented by the measurement model. In the case of the YCIS, we are testing whether the primary factor of counseling impact is dominant--expecting at the same time that there will be some indication of the two secondary factors of Insight and Change in the residuals. As mentioned above, we applied Linacre's (1998) approach to check for multidimensionality, which is integrated into the Winsteps software.

The results of this dimensionality check confirm that there is one dominant primary factor present in the YCIS. This primary factor represented by the measure itself explains 67.3% of the total variance while the next contrast explains only 7.7%. The factor loadings of the items on the primary factor range from .66 to .76. Thus, the Rasch measurement model requirements for a unidimensional scale are met in a satisfactory way.

However, the second contrast explains 23.5% of the residual variance making it worthwhile to explore it further. As can be seen by the loadings in Table 3, this contrast in the standardized residuals differentiates between the items of the two subscales. Items one through five (Insight) all load positively on this secondary factor while items six to ten (Change) all have negative loadings. In the next section we will provide some additional evidence for the presence of these secondary factors representing the two sub-scales.

Confirmatory Factor Analysis

The proposed hierarchical model of one primary general factor (Impact) and two secondary factors (Insight and Change) fit the empirical data well (CFI = .98; GFI = .96; RMSEA = .07) as can be seen in Table 4. The fit statistics for the one factor model indicate a worse fit (CFI = .89; GFI = .82; RMSEA = .15) and the Chi Square difference test is significant suggesting that the hierarchical G-factor model is the preferable model among these two. The standardized estimates of the factor loadings of the items on the two secondary factors were generally high ranging from .64 to .89 as can be seen in Table 5. The standardized loadings of Insight and Change on the general Impact factor are .86 and .94.

Overall these results confirm the proposed hierarchical factor structure. As mentioned earlier, the exploratory principal component factor analysis of the standardized residuals that was presented as part of the Rasch measurement section provided additional support for the proposed model.

Relationship with other variables

As can be seen in Table 6, the correlations of the YCIS with alliance (Y-TAQS) and with service satisfaction (Y-SSS) are both large (i.e., >.5) based on Cohen's standards. The correlation with treatment motivation (Y-MYTS) is medium (r = .35). Thus, as proposed, we further investigated the discrimination of the YCIS to the Y-TAQS and the Y-SSS. The pattern in the item-by-item correlation matrix provided evidence that the scores on the YCIS are related but distinct from the other two scales as hypothesized. The correlations of the items within the YCIS were on average clearly higher than those between the YCIS and the other two scales. In a couple of cases, the correlations of the items of related measures were close to the items of one of the subscales (e.g., Insight), but clearly lower than those of the other subscale (e.g., Change), thus providing further evidence for the validity of the subscale scores. The tables with these correlation matrices are available from the first author upon request.

Discussion

In this paper we described and evaluated the Youth Counseling Impact Scale, a new therapeutic process measure designed for frequent use in the mental health treatment of youths aged 11–18. The YCIS provides information about children and adolescent's impressions about the impact individual therapy sessions have on their thoughts, feelings, and behaviors. This measure provides a general Impact score as well as two subscale scores: Insight and Change. We used a mixed methods approach in the psychometric evaluation of this scale. This allowed us to use the strengths of each method as well as ameliorate its weaknesses by supplementing it with other methodological approaches. It also allowed us to triangulate certain results such as the hypothesized hierarchical G-model factor structure and thus provide stronger evidence for the validity of its use than we could have using just one of the methods.

Overall, the results are very positive. The YCIS is a short instrument with good psychometric properties and the characteristics of a measurement scale. The scores are approximately normally distributed in the intended population and the scale scores show good internal reliability, as evidenced by both the CTT and Rasch measurement approaches. The Rasch measurement analysis showed that the data fit the Rasch rating scale model reasonably well and, thus, demonstrates good scale characteristics. The Rasch analysis also highlighted two weaknesses of the scale. First, respondents seem to struggle with the differentiation of two of the answer categories. The expected improvement that a change in the answer categories would provide, however, is too minor to justify using a different rating scheme than that of other measures in the measurement battery. Second, this analysis showed the need to either drop or reword item ten that asks about change in school. We have changed the wording of this item in the most recent version of the YCIS. The overall fit indices of the Rasch model, however, are good in spite of these issues.

The confirmatory analysis provided support for the proposed hierarchical G-factor model with Impact as the primary general factor and Insight and Change as two secondary factors. Thus, the way we conceptualized counseling session impact theoretically, was resembled by the empirical data. Our theoretical model of impact was further strengthened by evidence, which showed that the scores on the YCIS had significant associations with measurement scores of related constructs such as alliance, treatment motivation, and service satisfaction, but also provided new and interesting information not captured by these other scales.

This first evaluation of the YCIS is very promising. The YCIS may be a helpful measure for researchers who want to explore how processes in individual counseling sessions may influence outcomes. It may also be useful for mental healthcare providers who want a brief tool to evaluate clients' perceptions of the positive impact of individual counseling sessions.

One possible shortcoming of the evaluated version of the YCIS is that it had no items asking about insights into to the youth's strengths, an important aspect of many therapeutic approaches. For the most recent version of the YCIS, we added an additional item that asks specifically about strengths. While we currently have no definitive psychometric information on the scale including this item, we believe that it is an important addition that will not alter the general character of the scale. Preliminary analyses of a new dataset including this item are very promising.

Further analyses are also needed to evaluate the measure's predictive validity as well its sensitivity to change. The YCIS is currently being used in the study led by Dr. Leonard Bickman at the Center for Evaluation and Program Improvement. In this study, using data from a large national mental health service provider, the YCIS is administered every other week for at least six months or until the youth is discharged. We expect over 1000 youths to participate in this study. Several other measures, all included in the Peabody Treatment Progress Battery, are

also being used in this longitudinal study which will allow us to assess the sensitivity to change as well as the scale's predictive power. We will also carefully investigate to what degree the scale could be shortened without losing its psychometric qualities and sensitivity to change. These findings will be presented in a separate paper. Researchers or providers who are interested in using the YCIS (or any other instruments in the Peabody Treatment Progress Battery) are encouraged to contact the first author about updates.

Acknowledgments

References

Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561–573.

Barak A, Bloch N. Factors related to perceived helpfulness in supporting highly distressed individuals through an online support chat. Cyber Psychology and Behavior 2006;9:60–68.

Bickman, L.; Riemer, M.; Lambert, EW.; Kelley, SD.; Breda, C.; Dew, SE.; Brannan, AM.; Vides de Andrade, AR., editors. Manual of the Peabody treatment progress battery. Nashville, TN: Vanderbilt University; 2007. [Electronic version].

Bollen, KA. Structural equations with latent variables. New York: Wiley; 1989.

Bond, TG.; Fox, CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahway, NJ: Earlbaum; 2001.

Bussell JR. Exploring the role of therapy process and outcome in interventions that target adolescent identity and intimacy. Dissertation Abstracts International: Section B: The Sciences and Engineering 2001;61(10-B):5553.

Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 1959;56:81–105. [PubMed: 13634291]

Clark LA, Watson D. Constructing validity: Basic issues in objective scale development. Psychological Assessment 1995;7:309–319.

Cohen J. A power primer. Psychological Bulletin 1992;112:155–159. [PubMed: 19565683]

Dunne A, Thompson W, Leitch R. Adolescent males' experience of the counseling process. Journal of Adolescence 2000;23:79–93. [PubMed: 10700374]

Elliott R, Wexler M. Measuring the impact of sessions in process-experiential therapy of depression: The Session Impacts Scale. Journal of Counseling Psychology 1994;41:166–174.

Embretson SE. The new rules of measurement. Psychological Assessment 1996;8:341–349.

Eugster SL, Wampold BE. Systematic effects of participant role on evaluation of the psychotherapy session. Journal of Consulting and Clinical Psychology 1996;64:1020–1028. [PubMed: 8916631]

Fitzpatrick MR, Iwakabe S, Stalikas A. Perspective divergence in the working alliance. Psychotherapy Research 2005;15:69–79.

Gregory MA, Leslie LA. Different lenses: Variations in clients' perception of family therapy by race and gender. Journal of Marital and Family Therapy 1996;22:239–251.

Holmes SE, Kivlighan DM. Comparison of therapeutic factors in group and individual treatment processes. Journal of Counseling Psychology 2000;47:478–484.

Ji PY. The role of attachment status in predicting longitudinal relationships between session-impact events and the working alliance within an adolescent client population. Dissertation Abstracts International: Section B: The Sciences and Engineering 2002;62:5966.

Karver MS, Handelsman JB, Fields S, Bickman L. A theoretical model of common process factors in youth and family therapy. Mental Health Services Research 2005;7:35–51. [PubMed: 15832692]

Kazdin AE, Nock MK. Delineating mechanisms of change in child and adolescent therapy: Methodological issues and research recommendations. Journal of Child Psychology and Psychiatry 2003;44:1116–1129. [PubMed: 14626454]

King R, Bambling M, Reid W, Thomas I. Telephone and online counselling for young people: A naturalistic comparison of session outcome, session impact and therapeutic alliance. Counselling & Psychotherapy Research 2006;6:175–181.

Kivlighan, DM., Jr; Angelone, EO. Helpee introversion, novice counselor intention use, and helpee-rated session impact. In: Hill, CE., editor. Helping skills: The empirical foundation. Washington, DC: American Psychological Association; 2001. p. 169-178.

Linacre JM. Detecting multidimensionality: Which residual data-types works best? Journal of Outcome Measurement 1998;12:266–283. [PubMed: 9711024]

Linacre, JM. Estimation methods for Rasch measures. In: Smith, EV., Jr; Smith, RM., editors. Introduction to Rasch Measurement. Maple Grove, MN: JAM Press; 2004. p. 25-47.

Linacre, JM.; Wright, BD. WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis. Chicago: MESA Press; 2000. [Computer software].

Lord, FM.; Novick, MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.

Mohr JJ, Gelso CJ, Hill CE. Client and counselor trainee attachment as predictors of session evaluation and countertransference behavior in first counseling sessions. Journal of Counseling Psychology 2005;52:298–309.

Nocita A, Stiles WB. Client introversion and counseling session impact. Journal of Counseling Psychology 1986;33:235–241.

Odell M, Quinn WH. Therapist and client behaviors in the first interview: Effects on session impact and treatment duration. Journal of Marital and Family Therapy 1998;24:369–388. [PubMed: 9677542]

Rasch, G. Probabilistic models for some intelligence and attainment tests. Rev. ed. Chicago: University of Chicago Press; 1980.

Reynolds S, Stiles WB, Barkham M, Shapiro DA, Hardy GE, Rees A. Acceleration of changes in session impact during contrasting time-limited psychotherapies. Journal of Consulting and Clinical Psychology 1996;64:577–586. [PubMed: 8698952]

Samstag LW, Batchelder ST, Muran JC, Safran JD, Winston A. Early identification of treatment failures in short-term psychotherapy: An assessment of therapeutic alliance and interpersonal behavior. Journal of Psychotherapy Practice and Research 1998;7:126–143. [PubMed: 9527957]

Smith, EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In: Smith, EV., Jr; Smith, RM., editors. Introduction to Rasch Measurement. Maple Grove, MN: JAM Press; 2004. p. 575-600.

Smith RM, Linacre JM, Smith EV. Guidelines for manuscripts. Journal of Applied Measurement 2003;4:198–204.

Stiles WB. Measurement of the impact of psychotherapy sessions. Journal of Consulting and Clinical Psychology 1980;48:176–185. [PubMed: 7365056]

Stiles WB, Reynolds S, Hardy GE, Rees A, Barkham M, Shapiro DA. Evaluation and description of psychotherapy sessions by clients using the Session Evaluation Questionnaire and the Session Impacts Scale. Journal of Counseling Psychology 1994;41:175–185.

Tryon GS. Session depth and smoothness in relation to the concept of engagement in counseling. Journal of Counseling Psychology 1990;37:248–253.

Wright BD, Linacre JM. Reasonable item mean-square fit values. Rasch Measurement Transactions 1994;8:370.

Wright, B.; Masters, G. Rating scale analysis: Rasch measurement. Chicago: MESA Press; 1982.

Wright, BD.; Mok, MMC. An overview of the family of Rasch Measurement Models. In: Smith, EV.; Smith, RM., editors. Introduction to Rasch Measurement. Maple Grove, MN: JAM Press; 2004. p. 1-24.
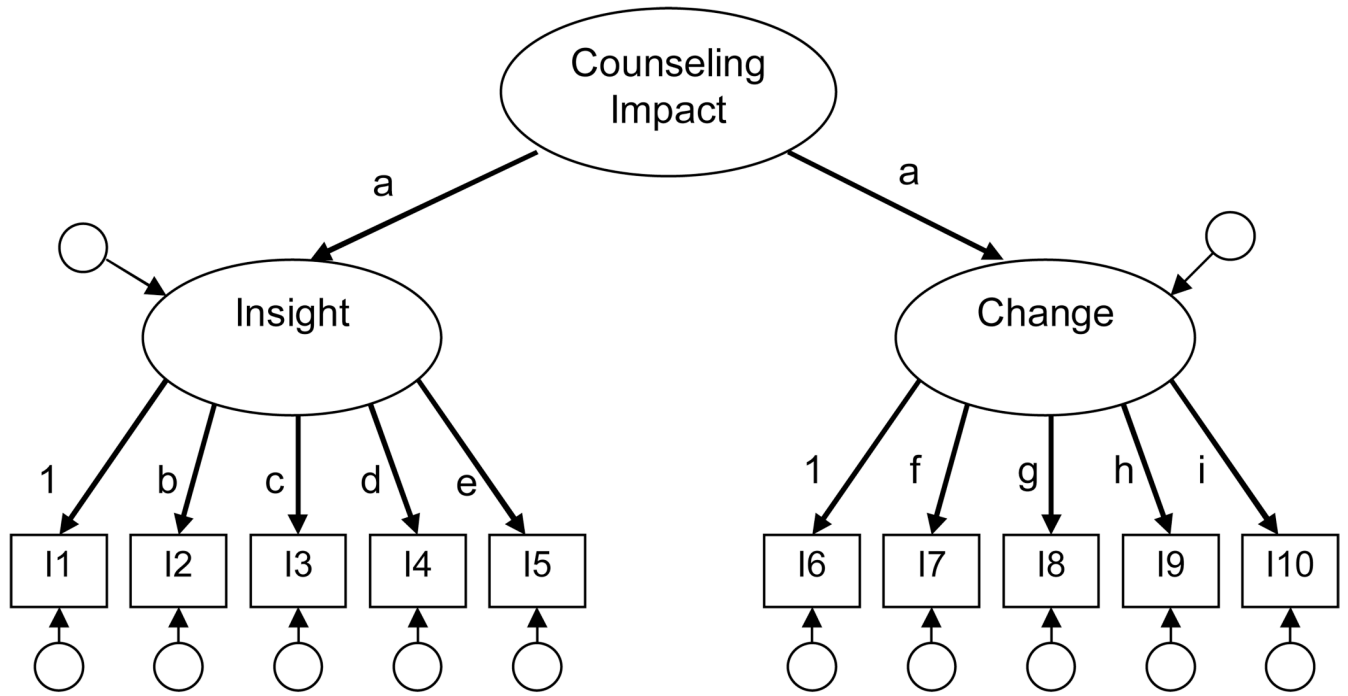
Figure 1.
Proposed Factor Structure for the YCIS

Category Probabilities: Modes - Structure Measures at Intersections
(1=Not At All, 2=Only A Little, 3=Somewhat, 4=Quite A Bit, 5=Totally)

```
P      ++---------+---------+---------+---------+---------+---------++
R  1.0 +                                                            +
O      |                                                            |
B      |1111                                                        |
A      |    111                                                    5|
B   .8 +       111                                            555  +
I      |          11                                         55     |
L      |            1                                       55      |
I      |           11                                      55       |
T   .6 +             1                                    55        +
Y      |              1                                  5          |
    .5 +               11                               55          +
O      |                 1         33333        44444 55            |
F   .4 +                  1    333       3334444     *444           +
       |                   *3              44333    55    444       |
R      |                  3 1            44      33 5        444     |
E      |                3*2   11    44         5*3            444    |
S   .2 +             2222**2  2222**4        55    33            444 +
P      |           22222 333      4441*22   555        333        44|
O      | 222222    333        444     11***2            333         |
N      |2    333333     44444     55555 111*22222        3333333     |
S   .0 +*************5555555555          11111**********************+
E      ++---------+---------+---------+---------+---------+---------++
        -3        -2        -1         0         1         2         3
```

Person [minus] Item MEASURE

Figure 2.
YCIS Item Characteristic Curve (generated by Winsteps)

```
              Person        |        Item

              <high impact>|<rare>
          4   ##########  +
                          |
                       .  |
                          |
                          |
                          |
          3           .## +
                          |
                       .  |
                          |
                     .##  |
                      .   |
                    .###  |
          2           .   +
                     .#   |
                    .###  |
                      .   |
                     .##  |
                    .###  |
                     .##  |
          1          .### +
                     .###  |
                     .###  |
                     .###  |
                     .## M|   Improved in school
                 .#######  |
                     ###   |   Understand others | Felt better about self | Used what learned
          0        .#####  +M  Learned about self | Tried things | Improved at home
                 .######   |   Understand feelings
                  .####    |   Understand problems | How to deal
                     .##   |
                    .###    |
                     .##    |
                     .##    |
         -1           #    +
                     .##    |
                      .     |
                      .     |
                      .     |
                      .     |
                      .     |
         -2                +
                      .     |
                            |
                      .     |
                            |
                            |
         -3          .   .  +
                            |
                            |
                            |
                            |
                            |
                            |
         -4          .#    +
              <low impact>|<frequent>

       "." = 1 person; "#" = 5 persons; M = Mean
```
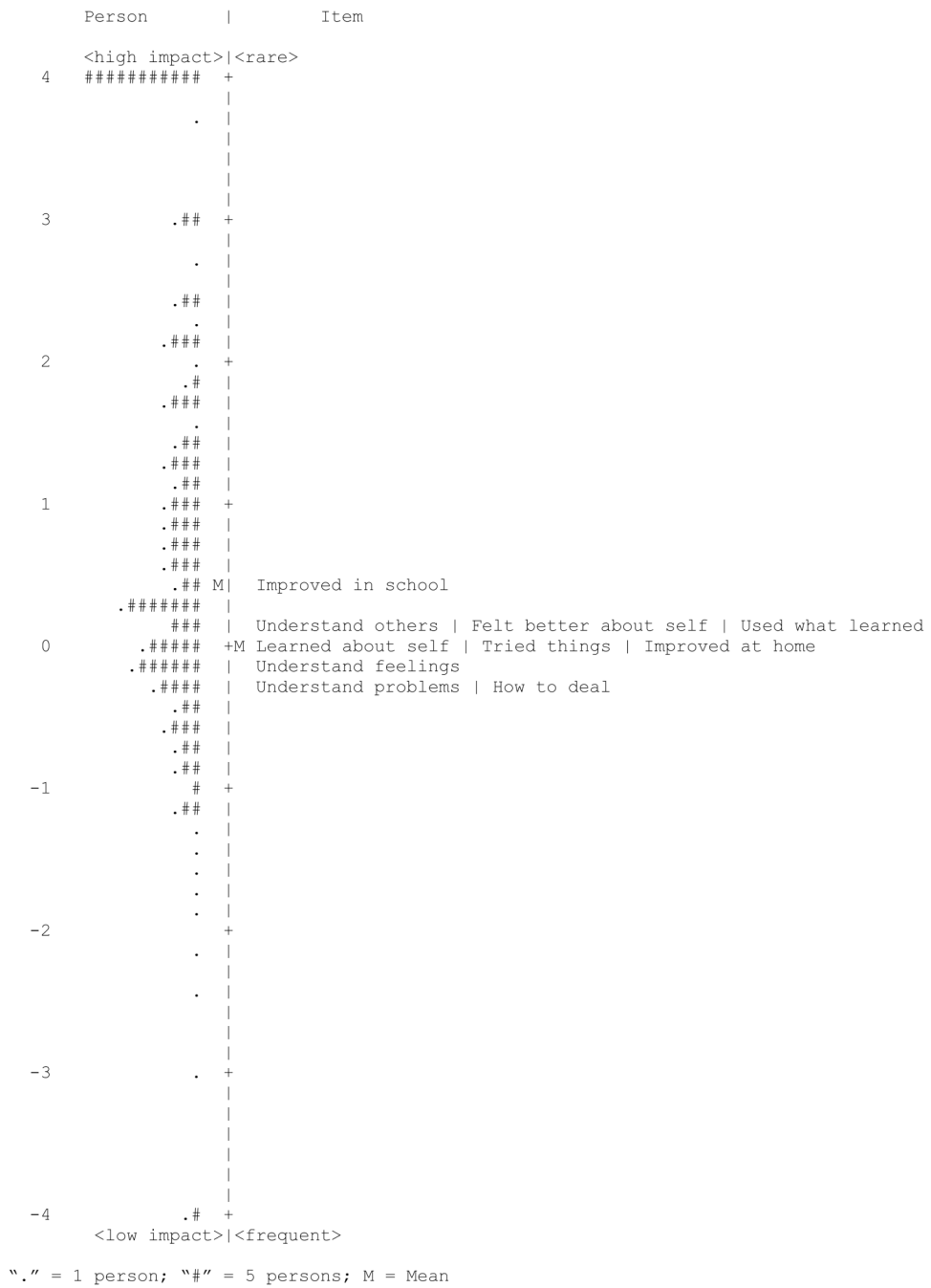
Figure 3.
Person Map of Items (generated by Winsteps)

Table 1

Results of Rasch Measurement Analysis: Person Indicators (N persons=442; N items=10)

	Mean Score	Measure	Infit		Outfit	
			MNSQ	ZSTD	MNSQ	ZSTD
Mean	3.59	.47	1.07	−.3	1.06	−.3
S.D.	1.0	1.16	.89	2.0	.87	2.0
Max	5	3.66	4.30	4.4	4.22	4.3
Min	1	−3.02	.03	−4.5	.03	−4.5

Person reliability = .80; MNSQ = Mean Square Value; ZSTD = Standardized Value

Table 2

Results of Rasch Measurement Analysis: Item Indicators (N persons=442; N items=10)

			Infit		Outfit	
	N	Measure	MNSQ	ZSTD	MNSQ	ZSTD
Mean	437.3	.00	1.01	−.2	1.06	.5
S.D.	-	.21	.29	4.0	.30	3.7
Max	442.0	.43	1.75	9.3	1.75	8.2
Min	408.0	−.34	.68	−5.2	.71	−4.4

Item reliability = .93; MNSQ = Mean Square Value; ZSTD = Standardized Value

Table 3

Factor Loadings for the Second Contrast from Principal Component Analysis of Standardized Residuals Using the Rasch Measurement Scores as the Primary Factor (N=442)

Item (Insight)	Factor Loading	Item (Change)	Factor Loading
(1) Learned about self	.42	(6) Tried things	−.42
(2) Understand others	.39	(7) Felt better about self	−.43
(3) Understand feelings	.59	(8) Used what learned	−.51
(4) Understand problems	.61	(9) Improved at home	−.50
(5) How to deal	.55	(10) Improved in school	−.35

Table 4

Goodness of Fit Indicators of the One Factor and the hierarchical G-Factor Models (N=450)

Model	X^2	df	X^2/df	X^2 diff	CFI	GFI	RMSEA
One Factor	371.55***	35	10.62		.89	.82	.15
G-Factor	103.36***	34	3.04	268.19***	.98	.96	.07

***
p<.001;

CFI=Comparative Fit Index, GFI=Goodness of Fit Index, RMSEA=Root Mean Square Error of Approximation

Table 5

Unstandardized Loadings (Standard Errors) and Standardized Loadings for the hierarchical G-Factor Confirmatory Model of Counseling Impact (N=450)

Item	Insight		Change		Impact	
	Unstand-ardized	Stand-ardized	Unstand-ardized	Stand-ardized	Unstand-ardized	Stand-ardized
(1) Learned about self	1.00 (--)	.78				
(2) Understand others	1.01 (.06)	.79				
(3) Understand feelings	1.13 (.05)	.89				
(4) Understand problems	1.10 (.05)	.85				
(5) How to deal	1.06 (.05)	.83				
(6) Tried things			1.00 (--)	.71		
(7) Felt better about self			1.16 (.07)	.82		
(8) Used what learned			1.23 (.07)	.87		
(9) Improved at home			1.07(.07)	.76		
(10) Improved in school			0.90(.07)	.64		
Insight					.67 (.03)	.86
Change					.67 (.03)	.94

Table 6

Correlations of the YCIS and its Subscales with Alliance, Motivation, and Service Satisfaction

	Impact		Insight		Change	
	r	(N)	r	(N)	r	(N)
Therapeutic Alliance	0.67***	(493)	0.63***	(493)	0.62***	(450)
Treatment Motivation	0.35***	(501)	0.38***	(501)	0.29***	(456)
Service Satisfaction	0.65***	(497)	0.60***	(494)	0.61***	(455)

***
p<.001;

r = Pearson correlation coefficient