Changes in Child Status During Behavioral Health Services in 2013:

Data from the Child and Adolescent Needs and Strengths Tool (CANS), Part 2, Domain Level Analysis

MassHealth Office of Behavioral Health
Boston, MA
August 30, 2016

Contents

Introduction 3
Domain Change versus Item Change 5
Domain scores and item scores 7
The dataset 9
Findings 11
Domain change scores 11
Testing domain change with the Reliable Change method 14
Summary and recommendations 17
Appendix 1: Item change averaged across domains 19
Appendix 2: How Domain Change scores relate to Item Change scores 21
The 4x4 table of pre / post scores on a CANS item 21
Interpreting the item change table 22
How item scores relate to domain scores 23
Example of how "percent resolved" relates to a domain change score 23
Appendix 3: Reliable Change Methodology 29
Appendix 4: Reliable Change Results with alpha = 0.05 32

Introduction

This is the second part of a two-part report, which together constitute the Commonwealth's first annual Standardized Analysis as described in MassHealth's *Plan for Ongoing CANS Data Analysis and Reporting*, issued April 29, 2015.[1] Part 1 of the Standardized Analysis report examined changes in *single* CANS items for children and youth in Intensive Care Coordination (ICC) and also for children and youth in In-Home Therapy (IHT). This Part 2 report looks at CANS items *grouped by domain*, and synthesizes findings and recommendations from both Part 1 and Part 2 analyses.

Part 1 reviewed briefly the function of the CANS tool in the MassHealth behavioral health system, and the item rating system that is the source of the CANS data. Please refer to Part 1 for those important contextual comments.

Much of the discussion in Part 2 is technical, related to issues of measurement and data analysis. Since the goal of this report is to report CANS findings in a way that is straightforward and comprehensible to all, technical discussion has been relegated, when possible, to appendices.

Findings from Part 1, focusing on item-level change, included the following:

- Changes varied considerably by item, even within domain.
- At the item level, children with more time in service tended to show more change (both increases and decreases in CANS scores). There are various explanations for this effect, one of which is that more treatment leads to more improvement or that children and families who experience more improvement (along with their clinicians), are motivated, as a result, to continue therapy longer. The data do not prove the cause of the changes. Indeed, probably more than one process underlies this trend in the data. Certainly it is important to look at factors that may lead families to end treatment without getting as much benefit as they might.
- For almost all CANS items, more children have decreases in ratings over time than have increases. In general, it is reasonable to believe that decreases in ratings over

[1] CANS is the acronym for the Child and Adolescent Needs and Strengths tool, developed by John S. Lyons PhD, copyright by the Praed Foundation, and modified for use by MassHealth. Part 1 of this report is available at www.mass.gov/eohhs/docs/masshealth/cbhi/changes-during-icc-and-iht-from-the-cans-part-1-dec2015.pdf

time signify an improvement in status. Many items had encouraging rates of resolution (item resolution up to 69%) within the timeframe of the study.[2]

- While *decreases* in CANS scores usually signify *improved* status, *increases* may not always indicate *worsened* status. Existing needs are not always known or disclosed by the family at the beginning of treatment. As the clinician gets to know the family, it is not uncommon to identify previously unrecognized needs. Thus an increased score may signify increased knowledge of the situation of the child and family. This probably explains an increase in ratings in the Caregiver Needs domain (e.g. family stress item, caregiver mental health item).
- An intensification of need, or emergence of new needs, may also occur as a consequence of child development. Risky behaviors tend to increase during adolescence, for example, as do new needs related to the transition to adulthood. This probably explains an increase in ratings in the Transition to Adulthood domain (e.g. Independent Living item, Financial Resources item).
- Items reflecting high risk behaviors tended to have high resolution rates, probably reflecting the high level of attention and intervention that is elicited by risky behaviors. Acute crises may also be, to some extent, self-limiting.
- By contrast, some issues that occur fairly often were not resolved as frequently. Examples include emotional control, hyperactivity / impulsivity, anxiety, and judgment. Resolution rates for these items were often around 25%. These comparatively low resolution rates raised questions about what clinical phenomena are being reflected in the CANS items (e.g., uncomplicated anxiety disorders versus complex trauma), and about what treatments are being directed to these conditions in ICC and IHT. While more information is needed to assess these outcomes, item outcomes do appear to provide guidance for quality improvement studies, particularly at the local program level.
- Identification of new concerns over time varied by item, but was often lower than one might expect if clinicians were conducting careful ongoing assessment and adjustment of CANS ratings.

---

[2] We defined *resolution* as a situation where a child with an initial rating of 3 or 2 on a CANS item had a subsequent rating of 1 or 0 on that item. Thus, a need which initially required intervention no longer required intervention, although it might warrant ongoing monitoring.

Domain Change versus Item Change

Examining change for all 66 items of the CANS is informative, but complicated. As CANS developer, Dr. John Lyons, has noted, "Single items can burden an analysis that seeks to make broad generalizations about a program or system."[3] To simplify the data, it makes sense to group together similar items. Grouping by domain is one way to do this.

CANS domains are groups of items that are conceptually related, as determined by the developer of the tool.[4] For example, the items in the Risk Behavior domain all relate, on their face, to risky behavior. CANS domains are best understood as a conceptual convenience for CANS users.

One could group items in other ways than by domain.[5] Psychometric measures, for example, often have subscales consisting of items grouped together based not on face content, but on statistical properties (usually on the extent to which they are correlated with one another). Although the current analysis groups items by domain (and does not drop items based on our preconceptions of whether they should change with intervention), we may discover over time that alternative approaches to grouping work better. We will return to this question in the discussion section of this report.

Thus, while Part 1 of this report focused on changes for single items, Part 2 groups items by domain, averaging item ratings across all the items of the domain. Since CANS items fall on a four-point scale from 0 to 3, the average of items will fall between 0 to 3 and will usually involve some figures after the decimal point (e.g., 1.78). It is conventional and convenient with CANS domain ratings to reduce the number of figures after the decimal by multiplying all scores by 10 (so 1.78 becomes 17.8, for example). We follow this convention in this report. Using this transformation, domain

---

[3] The CANS was explicitly designed by Dr. John Lyons according to "communimetric" as opposed to psychometric principles. Much of his book *Communimetrics* is devoted to comparison of the communimetric approach in contrast to the psychometric approach, as well as points of convergence. Lyons, J. S. (2009). *Communimetrics: A Communication Theory of Measurement in Human Service Settings.* New York: Springer. The quote is from *Communimetrics*, p 99.

[4] Although every jurisdiction that uses the CANS has the option to modify its content, the MassHealth CANS uses domains and item assignments created by the developer, Dr. John Lyons.

[5] Even when grouping by domain, Dr. Lyons suggests that certain items may be dropped from the analysis because they are not likely to change as a result of intervention.

scores fall on a range from 0 to 30. Change scores on a domain are similarly multiplied by 10.

For example, suppose a hypothetical domain A has ten items (typical for a CANS domain), and a child is rated as follows, initially and subsequently on each item (these are patterns one might easily find in a MassHealth CANS):

Item	Initial	Subsequent
item A1	0	0
item A2	0	0
item A3	1	1
item A4	0	0
item A5	1	1
item A6	2	2
item A7	2	1
item A8	3	2
item A9	0	0
item A10	2	2
Item average	1.1	0.9
Domain score	11	9

The initial item average is 1.1 and the initial domain score is 11.

Note that from the initial to the subsequent rating period, item 7 resolves from 2 to 1, and item 8 improves from 3 to 2, with no improvement of worsening on other items.

The new item average is 0.9 and the new domain score is 9. The domain change score for this individual will be 9 - 11 = -2 points.

Note that if a new problem previously rated as 0 (any of item 2, 4, or 9) had been identified in working with this family, and subsequently rated 2, this would eradicate gains made on items 7 and 8 and would result in a change score of 0 points. Averaging across groups of items thus allows newly identified problems to negate gains made on previously identified problems.

The initial domain score for *all* the children under consideration would be the average of their initial individual domain scores, and the subsequent domain scores for the group would be the average of their subsequent individual domain scores.[6] While the change score for the whole group of children could theoretically range from 0 to 30, the change score will typically be very much smaller than 30. For example, if every child in the group responded just like the child in the previous example, the group change score would be -2 points.

We have seen in Part 1 of this report that items behave differently from one another, even within domains. This is not at all surprising within the communimetric approach: "Given the design considerations of a communimetric tool, one would not expect different items to necessarily correlate with each other."[7] Averaging items that are very heterogeneous in their measurement behavior can be problematic from a measurement perspective, and ideally we would subject domain scores to statistical analysis (e.g., factor analysis) to confirm their psychometric validity.[8] Such analysis is beyond the scope of this report.

Domain scores and item scores

Since domains are composed of items, item change and domain change must be related. Appendix 2 describes in detail how the item change percentages reported in Appendix 1 relate to the domain change scores to be presented below. One lesson from Appendix 2 is that an item change frequency measure (such as percent resolved) always oversimplifies the behavior of an item.

Domain change scores also oversimplify the data, but in different ways from percentage of change on an item. Domain change scores incorporate all changes on each item (all the cells in the 4x4 table described in Appendix 2), but they average all positive and negative item changes. As discussed above, this can be problematic since (1) increased rating on an item has a different meaning from decreased ratings on the same item, and (2) behavior of items differs across items.

---

[6] See next section for definition of *initial* and *subsequent* ratings in this dataset.

[7] Lyons, *Communimetrics*, p. 69.

[8] "As soon as one seeks to use scale or dimension scores coming from a communimetric measure, the assumptions, considerations, and strategies that have arisen from psychometric theories become critical." Lyons, *Communimetrics*, p. 88.

Both methods of summarizing CANS changes discard some information, so their results should not be completely comparable (and will vary depending on their assumptions, such as how items are grouped or how the frequency of change is defined).

The dataset [9]

This report draws from complete CANS Five Through Twenty records entered into the CANS application on the Virtual gateway for dates of assessment between January 1, 2013 and December 31, 2014 (the "time window").[10]

The dataset was then filtered to retain only CANS records identified as produced in ICC or in IHT. For a child in ICC, all CANS records completed in ICC by a single provider organization during the time window were gathered together. For a child in IHT, all CANS records completed in IHT by a single provider organization during the time window were gathered together. Records entered by other organizations were not included because examination of CANS records suggests that reliability of CANS ratings is higher within a provider organization than across organizations. There was no requirement, however, that records be entered by the same individual Certified Assessor.

CANS item change scores were computed by taking the difference in ratings between an *initial* CANS and a *subsequent* CANS. The initial CANS was found by taking the first CANS for the child in the selected service in a nine month period (that is, no CANS were entered for the child by the provider organization for the selected service during the previous nine months). So for a child in ICC, the first ICC CANS record entered by the provider for the child in nine months was taken to be the initial record for the purpose of analysis.[11] For a child in ICC the subsequent CANS could be the third or fourth CANS in the set (counting the initial CANS as the first, and ordering the records chronologically). Since the CANS is ordinarily completed at three month intervals, the third CANS would ordinarily occur six months after the initial CANS, and the fourth CANS would ordinarily occur nine months after the initial CANS. For a child in IHT, we chose the second and third CANS for comparison to the initial CANS, representing time periods of approximately three months and six months. (We chose shorter

---

[9] This section repeats material from Part 1. CBHI is grateful for the help of Josh Twomey PhD of UMass Medical School for running the CANS data analyses.

[10] Not included in this dataset are CANS for children whose caregivers declined consent to enter the full CANS into the CANS application, children whose CANS record in the application was incomplete, or children who were under five on the date of assessment. Also not included are children whose providers did not comply with the MassHealth requirement to complete the CANS.

[11] Although each CANS record is marked by the clinician as "initial" or "reassessment", we saw evidence that these designations were occasionally inaccurate, so we chose the nine-month lookback procedure as our method of identifying the *initial* CANS for the purpose of this analysis.

comparison periods for IHT than for ICC because length of stay in IHT tends to be shorter than that in ICC.)

This resulted in four sets of change scores for each CANS items or CANS domains: change in ICC with 3 CANS, change in ICC with 4 CANS, change in IHT with 2 CANS, and change in IHT with 3 CANS. An individual child could occur in all four sets (if he or she was enrolled in ICC for at least twelve months as well as in IHT for at least nine months during the time window). We did not exclude a child's data if they were enrolled in both ICC and IHT (as often occurs) and also did not exclude a child's data based on *prior* enrollments.[12] A child enrolled in just one of the services could appear in two datasets if the enrollment was long enough (e.g. if the child had both a third and fourth CANS in ICC during the time window) or in one dataset (e.g. third but not fourth CANS in ICC) if the enrollment was shorter. A child whose enrollment was too short to produce the requisite number of CANS in the service would not appear at all.

The number of CANS records varies by service, time period, and item. Since the data reported here are calculated from all relevant CANS records, there is no sampling error, hence no reporting of confidence intervals (i.e. margin of error).

[12] Except that, as noted previously in our methodology for identifying the *initial* CANS record, we did exclude children with CANS for enrollments in the same service with the same provider in the previous nine months. Since children sometimes re-enroll in a service, it is possible that some children actually had more experience in the services than their CANS count would indicate.

Findings

We present the domain change data in two stages. First, we give the change scores as calculated on page 5 (with domain scores obtained by averaging across items and then multiplying by 10). Then, we show rates of "reliable change" according to a methodology explained in greater depth in Appendix 3. Understanding reliable change methodology helps to understand why the rates of change using this methodology are so low in this dataset.

Domain change scores

Tables 1 through 4 show the domain change scores for each service (IHT and ICC) across two intervals (approximately three and six months for IHT, approximately six and nine months for ICC).

Table 1
IHT first and second CANS (3 months)
n = 33157

Domain	Change
Life Domain Functioning	-0.24
Behavioral Emotional Needs	-0.45
Child Risk Behaviors	-0.25
Child Strengths	-0.43
Transition to Adulthood	0.20
Caregiver Resources and Needs	0.30

Although the domain change scores are small (equivalent to one item out of every twenty to forty items changing by one rating point) they are in the same direction seen in the single item analysis. Changes were in the direction of lower scores over time, except for the Transition to Adulthood and Caregiver Needs and Strengths domains. Reasons for this pattern were discussed above (page 4).

Table 2
IHT first and third CANS (six months)
n = 20995

Domain	Change
Life Domain Functioning	-0.34
Behavioral Emotional Needs	-0.54
Child Risk Behaviors	-0.27
Child Strengths	-0.68
Transition to Adulthood	0.42
Caregiver Resources and Needs	0.55

We also see here the pattern found in the single item analysis that children who have more time in the service tend to show more item change. As discussed previously, this is consistent with but does not prove the idea that longer service produces more effect.

Table 3
ICC first and third CANS (six months)
n = 11823

Domain	change
Life Domain Functioning	-0.25
Behavioral Emotional Needs	-0.34
Child Risk Behaviors	-0.20
Child Strengths	-0.68
Transition to Adulthood	0.48
Caregiver Resources and Needs	0.39

Effects are similar for ICC and IHT. Note that children in ICC often receive IHT also; we are not looking at the effects of these services in isolation.

Table 4
ICC first and fourth CANS (nine months)
n = 7841

domain	change
Life Domain Functioning	-0.31
Behavioral Emotional Needs	-0.43
Child Risk Behaviors	-0.25
Child Strengths	-0.85
Transition to Adulthood	0.49
Caregiver Resources and Needs	0.43

Again we see the relationship between time in service and size of changes.

We hypothesize several causes for why the changes are small, some of which were discussed in Part 1 and above.

First, as previously discussed, averaging items across a domain causes newly identified problems to negate problems in which improvement is occurring. This may look like no progress when in fact progress is occurring on two fronts, assessment and treatment.

Second, some items probably should not be expected to change with treatment. John Lyons suggests that certain items can be omitted *a priori* for this reason. Inclusion of items that cannot change will also attenuate effects. Items such as Medical / Physical or Developmental Disability could be candidates for omission. We did not make a priori guesses about excluding items for this analysis because we find CANS do not

necessarily follow our intuitions. We believe that items could be excluded as a result of further empirical analysis in the future, however.

Third, and very importantly, the CANS is not highly sensitive to downward change, especially in a population that does not focus on the most acute individuals with many ratings of 2 and 3. As discussed in part 1, it is relatively common for a rating of 3 to move to 2, but it is not easy for 2 to become 1. As long as an issue requires continued intervention, it is still rated 2 even if it is much improved. A child receiving treatment for anxiety, for example, might successfully develop skills and knowledge that reduce and manage the symptoms, and might move from weekly therapy to once a month. The steady CANS rating of 2, however, would not track this improvement. Similarly, a rating of 1 may continue for a long time with no active intervention, as long as there is a need to watch for recurrence of the issue.

Fourth, unreliable data can attenuate (weaken) effects. In Massachusetts, as in many other jurisdictions, many clinicians have resisted the CANS or the use of the CANS data system. If some clinicians are not invested in providing accurate data, then the poor data they contribute will tend to mask effects that appear in data entered by diligent clinicians. Unfortunately we have no way to identify good versus bad data. Improving ratings quality is part of the quality improvement challenge for CBHI going forward.

Fifth, treatment may be less effective than it should be. MassHealth's intensive case reviews, particularly in In-Home Therapy, have consistently shown variable skill and thoroughness in clinical work. MassHealth is currently engaged in a comprehensive multipart effort to strengthen quality in IHT.

Finally, the nature of the MassHealth population and the nature of the clinical work are such that changes over three to nine months may be significant but not dramatic, when all children are averaged together. Case reviews consistently reveal that many families believe services are helpful even when their children do not show marked symptomatic and functional improvement. Families nonetheless feel supported, experience newfound hope, and say they understand their children better. The CANS is not particularly sensitive to changes in these areas, which may presage later changes in other areas (such as more confident and calm parenting and more effective parental advocacy for the child). As briefly discussed in Part 1, child symptoms such as anxiety, depression, and impulsivity may often reflect complex trauma which is not quickly resolved, even with the best of treatment, especially in the context of ongoing multiple familial stressors. The change we are looking for may require long and steady support.

Testing domain change with the Reliable Change method

Ratings scales like the CANS are based on human judgment, and ratings are susceptible to error. Ratings error can be systematic (pushing scores consistently up, or consistently down), or it can be random in magnitude and direction (i.e., "noise"). In the case of random error, it is possible -- if one knows the amount of random error -- to predict how big a domain change would need to be to reach reasonable confidence that the change was not an artifact, a chance result of random error. This safeguard against being deceived by random error is called the Reliable Change methodology. This methodology is described in greater detail in Appendix 3.

Only changes greater than the Reliable Change Index (RCI) determined by this method are considered for analysis; any smaller change is considered untrustworthy. The Reliable Change Index for CANS items is typically around 2 to 4 points.[13] As described above, the average change for each domain in our data is in the neighborhood from 0.25 to 0.5 points, suggesting that changes for most individuals in this data set will not meet the Reliable Change threshold.

This is just what the data show when tested with the RCI.

Percentages of children or youth who experienced a reliable increase, a reliable decrease, or no reliable change (according to the Reliable Change methodology, with a 10% tolerance for mistaking an effect due simply to measurement error as a real change), are displayed in the Figure 1 on the following page.[14]

IHT2 and IHT3 indicate, respectively, children who had 2 CANS and 3 CANS with the same IHT provider organization during the time window, as described above in the section headed "The dataset". These individuals would typically have had approximately 3 months and 6 months of IHT between their initial and last CANS.

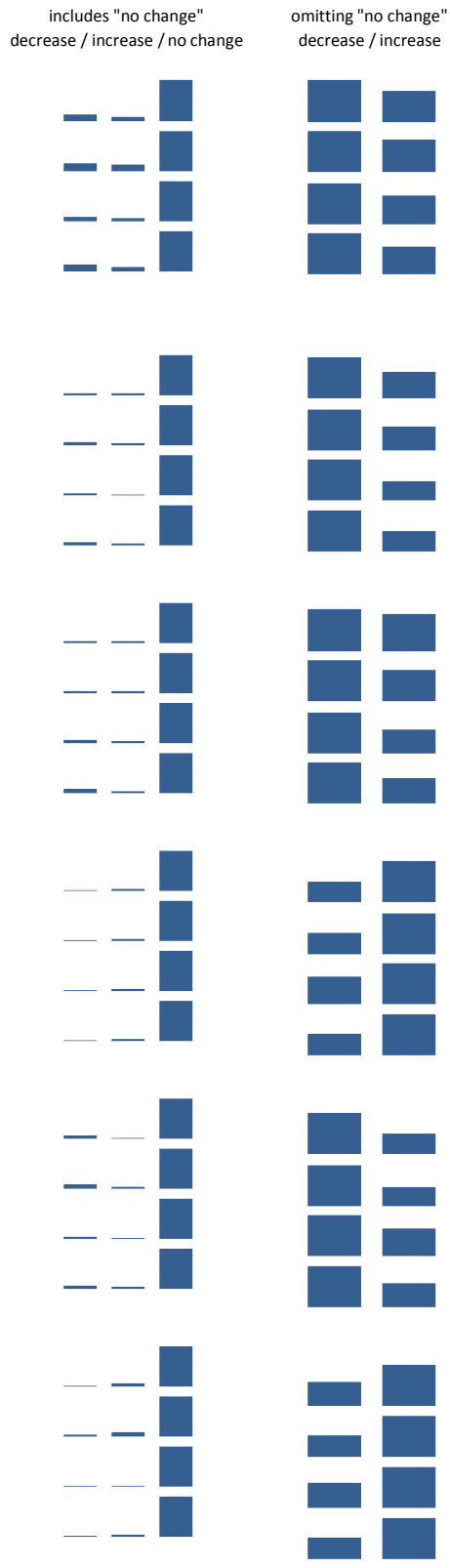
ICC3 and ICC4 similarly indicate, respectively, children who had 3 CANS and 4 CANS with the same IHT provider organization during the time window, as described above in the section headed "The dataset". These individuals would typically have had approximately 6 months and 9 months of IHT between their initial and last CANS.

---

[13] Lyons, *Communimetrics*, p. 121.

[14] A similar figure, differing only in setting a 5% tolerance instead of a 10% tolerance appears in Appendix 4, "Results with alpha = 0.05". Setting a smaller tolerance causes more individuals to fall into the "no reliable change" category, with a complementary reduction in the number identified as "reliable increase" and "reliable decrease".

Figure 1: Domain level changes, setting risk of "false change" at 10%



Domain	Population	% Decrease	% Increase	% No Change
LifeDomainFunc	ICC3	13.04	9.53	77.43
	ICC4	15.30	12.16	72.55
	IHT2	9.50	6.79	83.71
	IHT3	14.04	9.31	76.64

Domain	Population	% Decrease	% Increase	% No Change
Child Beh/Emo Needs	ICC3	6.25	4.08	89.67
	ICC4	7.95	4.68	87.37
	IHT2	5.71	2.61	91.67
	IHT3	7.79	3.91	88.30

Domain	Population	% Decrease	% Increase	% No Change
Child Risk Behaviors	ICC3	4.75	4.26	91.00
	ICC4	6.35	4.94	88.71
	IHT2	7.04	4.04	88.92
	IHT3	9.14	5.78	85.08

Domain	Population	% Decrease	% Increase	% No Change
Transition toAdulthood	ICC3	2.53	5.06	92.41
	ICC4	3.40	6.54	90.05
	IHT2	2.93	4.34	92.73
	IHT3	3.34	6.36	90.30

Domain	Population	% Decrease	% Increase	% No Change
Child Strengths	ICC3	6.92	3.55	89.53
	ICC4	9.29	4.48	86.24
	IHT2	5.68	3.85	90.47
	IHT3	8.44	4.88	86.68

Domain	Population	% Decrease	% Increase	% No Change
Caregiver Resources/Needs	ICC3	3.89	6.65	89.46
	ICC4	4.76	8.91	86.34
	IHT2	2.14	3.54	94.32
	IHT3	3.22	6.14	90.63

Since there are 4 service groups (IHT2, IHT3, ICC3, and ICC4), and 6 CANS domains (excluding Cultural Considerations, which informs treatment but is not a target of treatment), the table has 24 rows.

On the right of each row are two small bar plots. The first plot has three bars, representing the number of children whose domain score is classified as "reliable decrease", "reliable increase", and "no reliable change", respectively. The striking feature of this table and plot is the large number of individuals with "no reliable change", compared to a reliable change in either direction. "No change" ranges from a low of 77% to a high of 94%. Although the number of cases with a reliable change in either direction is relatively small, a two-bar plot on the far right shows how the percentages of "reliable decrease" and "reliable increase" compare (by omitting the "no change" bar).

The reader can observe that for most domains, decreases in domain scores occur more often than increases, but the reverse occurs for two domains: Transition to Adulthood and Caregiver Needs. Magnitudes vary somewhat by service group (ICC4, ICC3, IHT3, IHT2) but the overall patterns do not.

These findings are consistent with item-level analyses and are not surprising given the effect sizes reported in the previous section. Reasons why the magnitude of change is usually smaller than the RCI have already been discussed.

Summary and recommendations

Data presented in both parts of this report tell a similar story, involving these highlights:

- Changes in CANS scores tend to show a reduction in item ratings in most domains.[15] The two consistent exceptions are the domains of Transition to Adulthood and Caregiver Needs and Strengths. Increases are likely to occur with Transition to Adulthood as youth become older and the team becomes more aware of transitional issues. Increases are likely to occur with Caregiver items as families identify and disclose more caregiver issues over time.
- Changes tend to be greater for youth with more time in service. This may reflect the positive effect of longer engagement in services, which is certainly very plausible and clinically reasonable. On the other hand, the data do not prove causality and other factors may also be at play, such as a tendency for improvement to drive retention in services. In any case, the question of whether families are staying engaged for an appropriate length of time is an important one, better examined through case review methodologies such as the Massachusetts Practice Review and chart reviews routinely conducted by MassHealth's Managed Care Entities.
- Increase in an item rating does not necessarily have the reverse implication of decrease in that item. We believe that items with decreases tend to reflect perceived improvement rather than rating error. Items that increase, however, can likely reflect either worsening status (accurately rated), or increased identification of existing issues (a kind of measurement error that is entirely integral to the clinical process of increasing awareness). When an item (or domain) is averaged across individuals this distinction is lost, so positive clinical improvement and increased awareness can appear as no net progress.
- Changes for individuals in IHT and ICC are not notably different. This is hardly surprising given that most youth in ICC also receive IHT or other clinical service, and that many youth in IHT also receive ICC at some point. The reported data do not tell us about the impact of services in isolation.

[15] The Cultural Considerations Domain was not reported as this domain informs treatment but is not a target of treatment.

- Item-level data showed highest resolution rates for risk issues. This is a positive finding given the need to deal effectively with risky behavior. Resolution rates for items that may reflect complex trauma and long-term social support issues were much lower. This may reflect a reality of working with traumatized children and families who continue to function in the face of multiple stressors associated, in particular, with poverty. It is important to take a long view in assessing the impact of services, particularly in supporting caregivers, and not to focus solely on the child's behavior or on short term outcomes (such as three to nine month as in the current report). On the other hand, it is also important to ask whether children are getting the most effective services possible in order to deal with their complex challenges and whether clinicians and provider agencies have the trauma-informed training, strong supervision and clinical consultation, and other practices that will provide strongest clinical outcomes.
- Questions about the reliability of CANS ratings need to be further addressed by MassHealth so that practitioners will trust and use the data. Higher reliability will also clarify aggregate outcomes (such as reported in this document) by reducing noise in the data. The new CANS training and certification program launched in May 2016 are designed to address this need and will be followed by other CANS quality improvement initiatives currently in planning for FY17 and subsequent years.

In general the item-level data tend to be more interpretable than the domain-level data. While combining items into item groups has some appeal, the existing domains may not be the best groups for this purpose. Furthermore, the use of the RCI methodology is misplaced in this context and should be abandoned. Therefore, although the Commonwealth has agreed to produce annual reports on domain change, this activity would be of dubious value and should be reconsidered.

From a MassHealth perspective, the best use of CANS data in the foreseeable future will not be to evaluate system impact, but to improve the assessment and treatment of individuals, by focusing on quality improvement initiatives that integrate CANS practice into broader practice issues in services such as IHT. Both at the individual and provider agency level, CANS data help to understand the challenges that children and families face, and to shape programs to support them in the most effective way. MassHealth is now implementing a comprehensive initiative to strengthen IHT. While CANS data may be of some use in evaluating the impact of this initiative, more useful data will come from individual level reviews.

Appendix 1: Item change averaged across domains[16]

Domain	Population	Decreased / Improved	Increased / Worsened	Resolved	Newly IDd
LifeDomainFunc	ICC3	26%	13%	30%	8%
LifeDomainFunc	ICC4	31%	15%	37%	9%
LifeDomainFunc	IHT2	25%	10%	29%	5%
LifeDomainFunc	IHT3	32%	13%	39%	6%
Child BehEmo Needs	ICC3	26%	10%	31%	6%
Child BehEmo Needs	ICC4	31%	12%	38%	8%
Child BehEmo Needs	IHT2	26%	9%	30%	4%
Child BehEmo Needs	IHT3	33%	12%	40%	6%
Child Risk Behaviors	ICC3	34%	7%	49%	3%
Child Risk Behaviors	ICC4	42%	9%	60%	4%
Child Risk Behaviors	IHT2	34%	5%	48%	1%
Child Risk Behaviors	IHT3	44%	7%	62%	2%
Caregiver Resources/Needs	ICC3	19%	11%	24%	6%
Caregiver Resources/Needs	ICC4	24%	14%	30%	7%
Caregiver Resources/Needs	IHT2	20%	10%	28%	4%
Caregiver Resources/Needs	IHT3	24%	14%	36%	6%

[16] Adapted from Part 1 of this report, pages 9 - 11.

This table is explained more fully in Part 1 of this report. It presents information on change by domain, averaging over items in each domain. The "resolved" column is highlighted because it was the focus of discussion in Part 1.

It presents the percentage of children in each services group (IHT or ICC with different numbers of CANS recorded during the time window) who follow into one of three classes on CANS items:

- Those with decreased ratings ("Decreased / Improved") -- these children initially have a rating of 1, 2, or 3, which subsequently decreases by at least 1 point.
- Those with increased ratings ("Increased / Worsened") ") -- these children initially have a rating of 0, 1, or 2, which subsequently increases by at least 1 point.
- Those for whom new problems were resolved ("Resolved") -- these children initially have a rating of 2 or 3, which subsequently decreases to 1 or 0. This column was highlighted in Part 1 since it was the principal focus of discussion.
- Those for whom new problems were identified ("Newly IDd") -- these children initially have a rating of 0, which subsequently increases to 2 or 3.

Appendix 2: How Domain Change scores relate to Item Change scores

Since domains are composed of items, item change and domain change must be related. This section explains how measurement of domain change relates to measurement of item change.

It will help to first look closely at how we measure item change, and then translate this into measurement of domain change. Item change is most completely understood by a 4x4 table that contains all the information.

The 4x4 table of pre / post scores on a CANS item

		End (subsequent CANS) rating				
Start (initial CANS) rating		0	1	2	3	total start
	0	n_{00}	n_{01}	n_{02}	n_{03}	n_{0start} = sum of cells to the left
	1	n_{10}	n_{11}	n_{12}	n_{13}	n_{1start} = sum of cells to the left
	2	n_{20}	n_{21}	n_{22}	n_{23}	n_{2start} = sum of cells to the left
	3	n_{30}	n_{31}	n_{32}	n_{33}	n_{3start} = sum of cells to the left
	total sub-sequent	$n_{0subsequent}$ = sum of cells above	$n_{1subsequent}$ = sum of cells above	$n_{2subsequent}$ = sum of cells above	$n_{3subsequent}$ = sum of cells above	n_{total} = sum of all cells

In this case the table shows data for *one item* for *multiple children*.

The table counts every child in terms of where the child started and ended on the item. There are 4 possible starting ratings [0 to 3] and 4 possible ending ratings [0 to 3], and therefore 16 possible combinations of starting and ending ratings.

So,

- n_{00} is the number of children who started with a 0 and ended with a 0 (the first number of the subscript denotes the row, and the second number of the subscript denotes the column),
- n_{01} is the number of children who started with a 0 and ended with a 1,
- n_{10} is the number of children who started with a 1 and ended with a 0,

and so on.

Interpreting the item change table

Cells *on the diagonal* (containing n_{00}, n_{11}, n_{22}, n_{33}) contain the counts of children whose ratings (0, 1, 2, 3, respectively), did not change.

Cells *above the diagonal* (cells shaded red) contain the counts of children whose ratings increased. This could happen as a result of worsened status, or improved information about previously existing problems. The further from the diagonal, the more change occurred.

Cells *below the diagonal* (cells shaded blue) contain the counts of children whose ratings decreased. Change in this direction is likely to indicate improved status. Again, the further from the diagonal, the more change occurred.

The *sum of all cells* (lower right corner) is the total number of children in the data set for the item.

The *right-hand margin* of the table contains the count of children with *start* scores of 0, 1, 2, 3 respectively.

The *bottom margin* of the table contains the count of children with *subsequent* scores of 0, 1, 2, 3 respectively.

While the 4x4 table is used here to show change in *one item* for *multiple children,* it can also show change for *multiple items* for *one child*, as we will see in a subsequent example. And, most powerfully, we will see that it can also be used to show change in *multiple items* for *multiple children*.

How item scores relate to domain scores

Imagine a domain consisting of just the single item in the table.

- The *starting domain score* would be the average of the starting item scores across all children, multiplied (according to the convention described above) by 10. We can get this from the right hand margin of the table, weighting each value by multiplying by the starting item score. The weighting of the cell counts recognizes the fact that big changes in item ratings affect the domain score more than small changes.

That is,

Starting domain score = $[(0 \times n_{0start}) + (1 \times n_{1start}) + (2 \times n_{2start}) + (3 \times n_{3start})] \times 10$

- Similarly, the *subsequent domain score* would be the average of the subsequent item scores across all children, multiplied (according to the convention described above) by 10. We get this from the bottom margin of the table, weighting each value by multiplying by the subsequent item score.

Subsequent domain score = $[(0 \times n_{0subsequent}) + (1 \times n_{1subsequent}) + (2 \times n_{2subsequent}) + (3 \times n_{3subsequent})] \times 10$

- Finally, the domain change score is the subsequent domain score minus the starting domain score.

Example of how "percent resolved" relates to a domain change score

We repeat here the item change example from page 6, where domain A has 10 items as shown on the next page:

Item	Initial	Subsequent
item A1	0	0
item A2	0	0
item A3	1	1
item A4	0	0
item A5	1	1
item A6	2	2
item A7	2	1
item A8	3	2

item A9	0	0
item A10	2	2
Item average	1.1	0.9
Domain score	11	9

This table provides full information about the initial and subsequent states of the CANS item for one child for this domain, allowing us to complete the 4x4 table as follows:

	End (subsequent CANS) rating					
Start (initial CANS) rating		0	1	2	3	Total start
	0	4				4
	1		2			2
	2		1	2		3
	3			1		1
	total sub-sequent	4	3	3	0	10

In this case the table shows counts for *one child* with *multiple items*. So the total in the lower right corner is a count of total items for one child.

Since this 4x4 table shows counts from just one child, many cells are empty. But even with one instance, we see a common pattern: many of the filled cells are on the diagonal, indication no change from start to subsequent CANS. More off-diagonal changes occur in the blue regions (decreased ratings) than in the red regions (increased ratings). No cells are filled in the red regions, since no item had an increased score (no new needs emerged during this interval from initial to subsequent CANS).

In Part 1 (page 6) of this report we defined four ways of summarizing change for a single item, which can be computed from the 4x4 table as indicated:

- If a child initially has a score of 3 or 2, which subsequently becomes a 0 or 1, we say their need on the item is "*Resolved*". This is the most common scenario for needs that are successfully addressed by a service.
- If a child's score decreases then we say their status on that item is "*Decreased / Improved*". Most improvement occurs when ratings of 2 are reduced to 1; large numbers of 1 ratings do not improve due to the design of the CANS. For this reason the rate of items Resolved is actually usually higher than the rate of items Improved.
- If a child's score increases then we say their status on that item is "*Increased/Worsened*". This can reflect an actual deterioration in status, or the acquisition of more accurate information about the severity of a need. Deterioration in status may occur for reasons related to external stressors or developmental factors, even when effective services are in place.
- If a child initially has a 0 on an item, which subsequently becomes a 2 or 3, we say a need on that item is "*Newly Identified*". We expect new needs to be identified fairly frequently during the course of services. This seems especially likely for items that we believe tend to be underrated, such as youth substance use and parental substance abuse and mental illness.

Each of these four methods provides a single percentage to summarize a complex pattern of change. For clarity we refer to these methods (and others that could be constructed along similar lines) "*simple percentage measures of CANS change*".

In this example, the child initially has four items with a score of 2 or 3 (items 6, 7, 8, and 10). Item 7 subsequently becomes a 1. So the number of items resolved is 1/4 or 25%. The fact that item 8 decreases from 3 to 2 is irrelevant in this case, since an rated 2 is not *resolved* (although that item would count if we were looking at items *improved*).

Simple percentage measures of CANS change, such as *resolved* or *improved* (as reported in Part 1 of this report) may indeed provide useful simplification, but they all discard, in one way or another, some potentially important part of the information in the 4x4 table. Each of these simple percentage measures reduces the 4x4 table, which displays 16 counts (corresponding in statistical language to *15 degrees of freedom*, or independently variable quantities) to just *1 degree of freedom* as expressed in a measure such as "percent of items resolved". While wisely discarding data in order to achieve simplification is the foundational accomplishment of descriptive statistics (for example, in arriving at means or standard deviations) it is always important to know: what information is being discarded? Does the simplification clarify or confound?

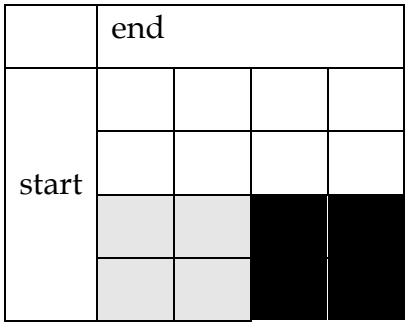
As we will see below, each simple percentage measure of CANS change ignores some cells of the 4x4 table altogether. One lesson from this cycle of reporting is that consumers of CANS change data for groups of children at the items level should learn to look first at the 4x4 table for the item, rather than immediately resorting to a simple percentage measure.

In this case, if every child were like the child in the example, and every item of the domain were like the item in the example, we would have a *resolved* rate of 25% for the group of children, along with a domain change score of -2 for the group. If the items did not all behave the same (and the do not) then the domain change score would be the average of the change scores for the items. In general, when more items have higher resolution rates, then domain changes will tend to be more favorable (in the negative direction). Similarly, when more items have higher improvement rates then domain changes will tend to be more favorable. When more items have lower rates of worsening, domain changes will tend to be more favorable.

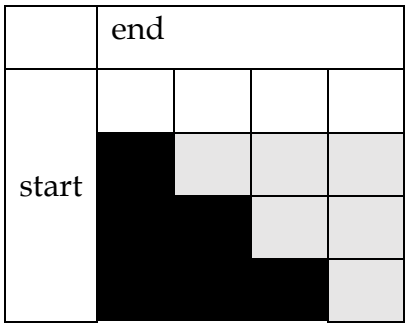
We have now moved to using the 4x4 table to show changes on *multiple items* for *multiple children*. In a table of this kind, the lower right hand cell counts the number of children times the number of items times the number of children. This illustrates the most valuable use of the 4x4 table for understanding CANS change in future work.

Using the 4x4 table, we show these four patterns as follows, where the gray cells and black cells are the denominator and the gray cells are the numerator, and whatever is in the unshaded cells is irrelevant:

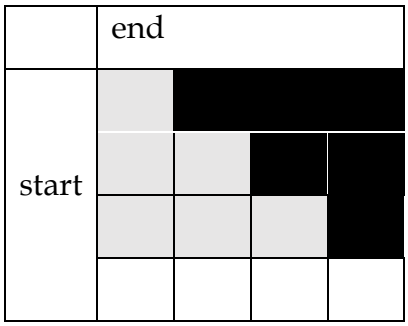
Resolved:



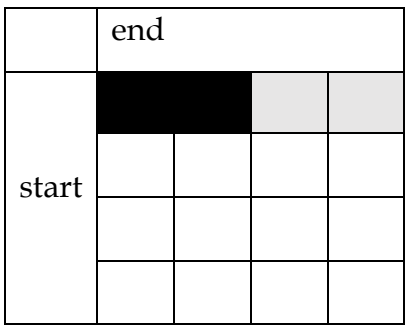
Decreased / Improved:



Increased/Worsened:



Newly Identified:



For those who wish to use simple percentage measures of CANS change, these patterns help to explain which information about CANS change is considered, and which information is discarded by the measure.

More generally, these examples illustrate the value of the 4x4 table in understanding the full picture of multiple items assessing multiple children over time.

Appendix 3: Reliable Change Methodology

If the characteristics of error in a scale are known, it should be possible to determine the amount of change that would be needed to ensure that the change was not due to random error alone.

Psychometric analysis generally imposes several requirements on a measurement process: that observations are independent of one another, and that error in the score is random and normally distributed (that is, follows a "bell curve"). If these requirements are met, and the reliability of the scale has been measured empirically, then the amount of measurement error in the scale can be calculated from the reliability. If one knows the amount of measurement error, and one determines the level of risk one is willing to tolerate of getting a false positive (a change being due to measurement error alone), one can derive the Reliable Change Index, which is the amount of change that is necessary for one to consider a change real, rather than an artifact due to error.[17]

Using the Reliable Change (RC) methodology with the CANS poses two challenges: it requires that the scales meet the psychometric criteria described above, and that the amount of error (that is, that the reliability of the domain score) must be known.

Regarding the first challenge, there is reason to doubt that CANS domain scores meet psychometric criteria. Clinicians updating the CANS are rarely starting from a blank slate; in most cases they are looking at previous CANS ratings for the child while they enter the update. Furthermore, it is questionable that measurement error in CANS ratings (and the domain scales that are built from them) are random and normally distributed. Based on experience with the CANS rating system and with accumulated CANS data, we believe that errors in initial CANS needs ratings are more likely to be underestimates than overestimates, and that ratings of an individual are likely to become more accurate over time as more information becomes available.[18]

---

[17] Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12–19. The level of risk is denoted as "alpha" by Jacobson and Truax and often set at 5%. Alpha of 5% means a one-in-twenty chance of accepting an artefactual change as real. In our analysis we use an alpha level of 10% but report results for alpha = 0.05 in an appendix.

[18] As discussed in Part 1 of this report,

Regarding the second challenge, there are several ways of measuring the reliability of a measure.[19] The method chosen for the RCI methodology should reflect the actual use of the CANS. While John Lyons sometimes recommends that the average inter-rater reliability of examinees on each domain of the CANS certification exam (their agreement with an expertly determined "correct" rating) be used to calculate the reliability of the domain, this estimate of reliability is actually quite different from the actual use of the CANS in the MassHealth system:

- **Reliabilities based on the certification examination** reflect the performance of an individual in a high-stakes examination setting, using data from a written vignette, compared to an expert rating based upon the same written vignette.
- **In practice, longitudinal CANS ratings as analyzed in this report** reflect the performance of one or more individuals NOT in a high-stakes examination setting, using rich information sources (not a limited vignette) and operating within the same provider agency and medical record system, and entering follow-up CANS with knowledge of previous CANS (rating events not independent.

These two reliability contexts differ greatly. Rater motivation is much higher in the first situation compared to the second, which would tend to increase rater accuracy. But rater knowledge of the child is likely to be higher in the second situation, and successive raters are more likely to agree even if they are not rating accurately (because the raters are actually the same person, or because they are working from the same medical record).

We are not confident that certification exams provide a model for actual CANS practice, or that certification data can be used to provide reliability estimates for actual CANS practice. For this reason, for this report we use reliability figures from a paper by John Lyons, based on field audits of CANS reliability.[20]

---

[19] For example: inter-rater, test-retest, parallel forms, internal consistency. Other versions have also been defined. See, for example, "Types of Reliability" in the Research Methods Database at http://www.socialresearchmethods.net/kb/reltypes.php, retrieved December 28, 2015.

[20] Anderson, R. L., Lyons, J. S., Giles, D. M., Price, J. A., & Estle, G. (2003). Reliability of the Child and Adolescent Needs and Strengths-Mental Health (CANS-MH) Scale. *Journal of Child and Family Studies*, 12(3), 279–289.

Although the version of the CANS used in the paper is not identical to that used by MassHealth, it includes the same domains and it is likely that their reliability is comparable. Reliabilities range from 0 (no reliability) to 1 (perfect reliability):

- Child Emotional and Behavioral Problems = 0.72
- Child risk behaviors = 0.76
- Life domain functioning = 0.85
- Caregiver = 0.75
- Child Strengths = 0.77
- Transition to adulthood = 0.75

Appendix 4: Reliable Change Results with alpha = 0.05

Figure appears on next page. For explanation, please see “Findings”.

Note that making alpha smaller will usually increase the number of youth with “no reliable change” while decreasing the number of youth in the categories “reliable increase” and “reliable decrease”. Occasionally, however, changing alpha may have no effect. This is not an error, but reflects a situation where no individuals had scores that would lead to reclassification with a change in alpha.

Domain level changes, setting risk of "false change" at 5%

Domain	Population	% Decrease	% Increase	% No Change
LifeDomainFunc	ICC3	7.94	6.34	85.72
	ICC4	10.15	7.62	82.23
	IHT2	5.95	4.13	89.92
	IHT3	9.14	5.25	85.61

Domain	Population	% Decrease	% Increase	% No Change
Child Beh/Emo Needs	ICC3	6.25	4.08	89.67
	ICC4	7.95	4.68	87.37
	IHT2	3.40	1.73	94.87
	IHT3	4.61	2.51	92.88

Domain	Population	% Decrease	% Increase	% No Change
Child Risk Behaviors	ICC3	4.75	4.26	91.00
	ICC4	6.35	4.94	88.71
	IHT2	4.07	2.18	93.75
	IHT3	5.90	3.29	90.81

Domain	Population	% Decrease	% Increase	% No Change
Transition toAdulthood	ICC3	1.69	3.71	94.60
	ICC4	2.09	4.97	92.93
	IHT2	2.14	2.69	95.17
	IHT3	2.40	4.17	93.43

Domain	Population	% Decrease	% Increase	% No Change
Child Strengths	ICC3	6.92	3.55	89.53
	ICC4	9.29	4.48	86.24
	IHT2	3.93	2.69	93.38
	IHT3	6.05	3.46	90.49

Domain	Population	% Decrease	% Increase	% No Change
Caregiver Resources/Needs	ICC3	2.40	4.39	93.21
	ICC4	3.20	5.78	91.03
	IHT2	1.47	2.32	96.21
	IHT3	2.09	4.43	93.48