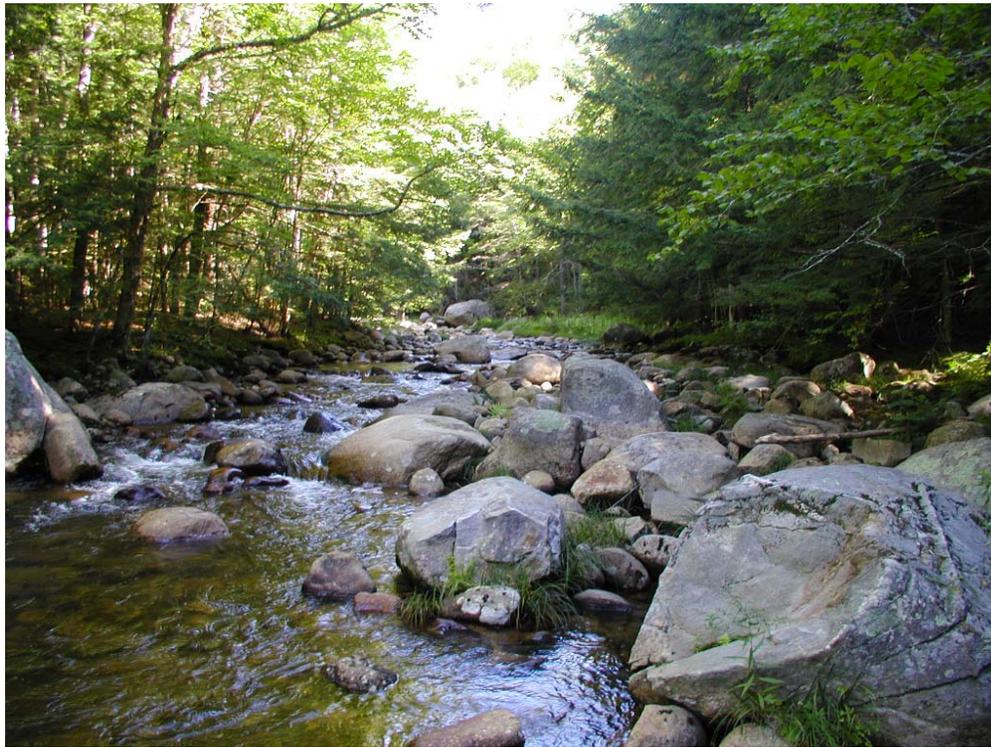
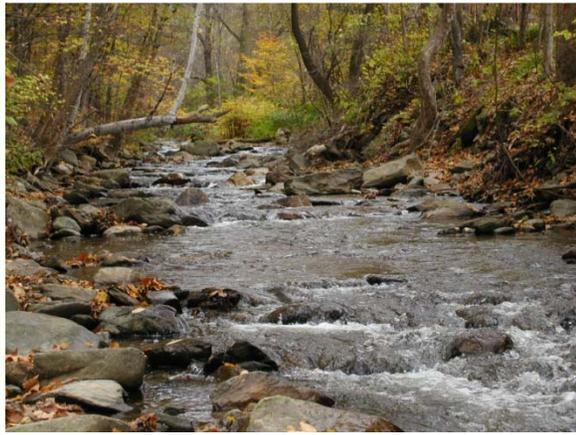


National Wadeable Stream Assessment: A comparison of eastern assessment outcomes

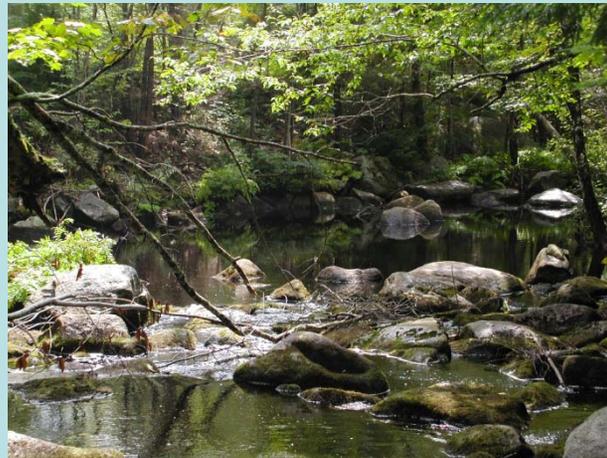


New England States participation in the WSA

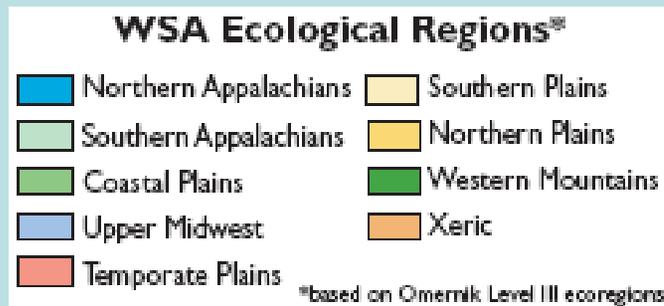
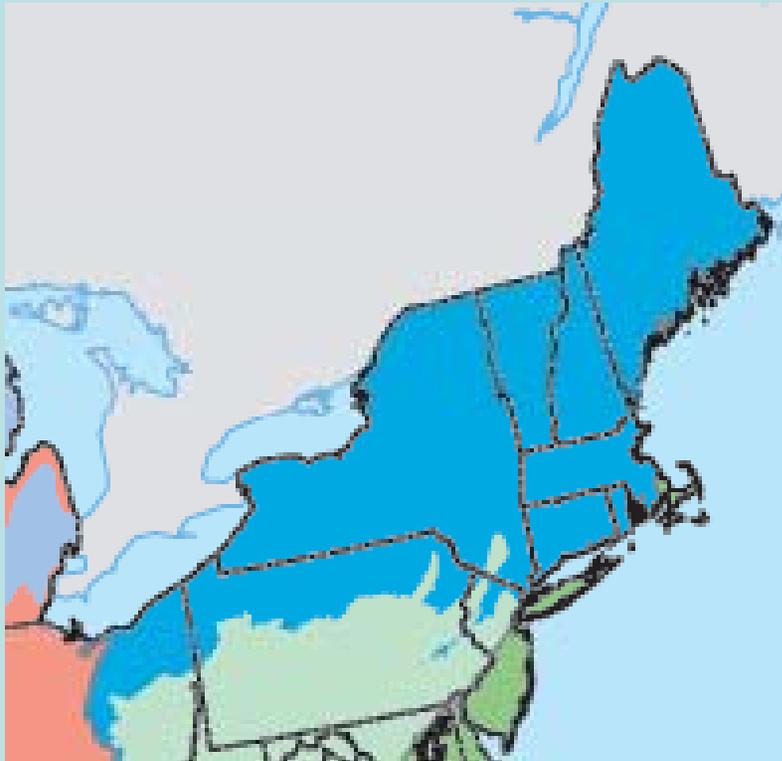
- NH DES, VT DEC, CT DEP, ME DEP established cooperative agreement with New England Interstate Water Pollution Commission (NEIWPCC)
- Decision to complete methods comparability data collection at national sites
- CT, ME, NH, and VT
- Macroinvertebrate samples collected at each site using up to 6 methods (CT, ME, NH, VT, WSA, NEWS)
- Laboratory sample processing done using each respective entity's method
- Final assessment status (Impaired, Not Impaired) at all locations using state biocriteria (CT, NH, VT) or regional thresholds (WSA)

Outline

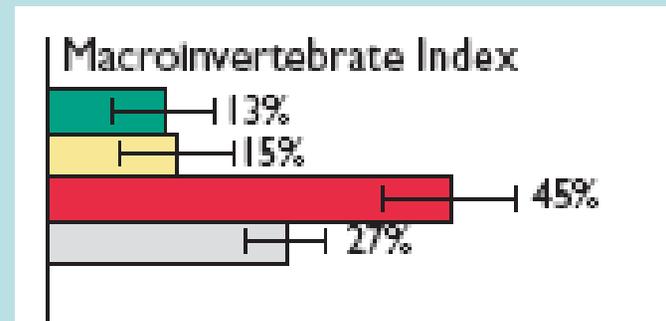
- Overview of NAP WSA results
- Summary of NE participating states sampling & condition tools
- Comparison of assessment outcomes
- Comparison of MMI scores for applicable condition tools
- Prediction of MMI scores



WSA Ecoregional Findings for Northern Appalachians (NAP)



- 85 sites
- 97,913 stream miles
- 13% good, 15% fair, 45% poor, 27% unassessed
- 45% high P, 45% high N
- 20% high riparian disturbance, 26% poor vegetative cover
- 29% “poor” instream sediments



For Macroinvertebrate Index:
 Good  Fair  Poor  Not Assessed

New States WSA Comparability Study Goal

?

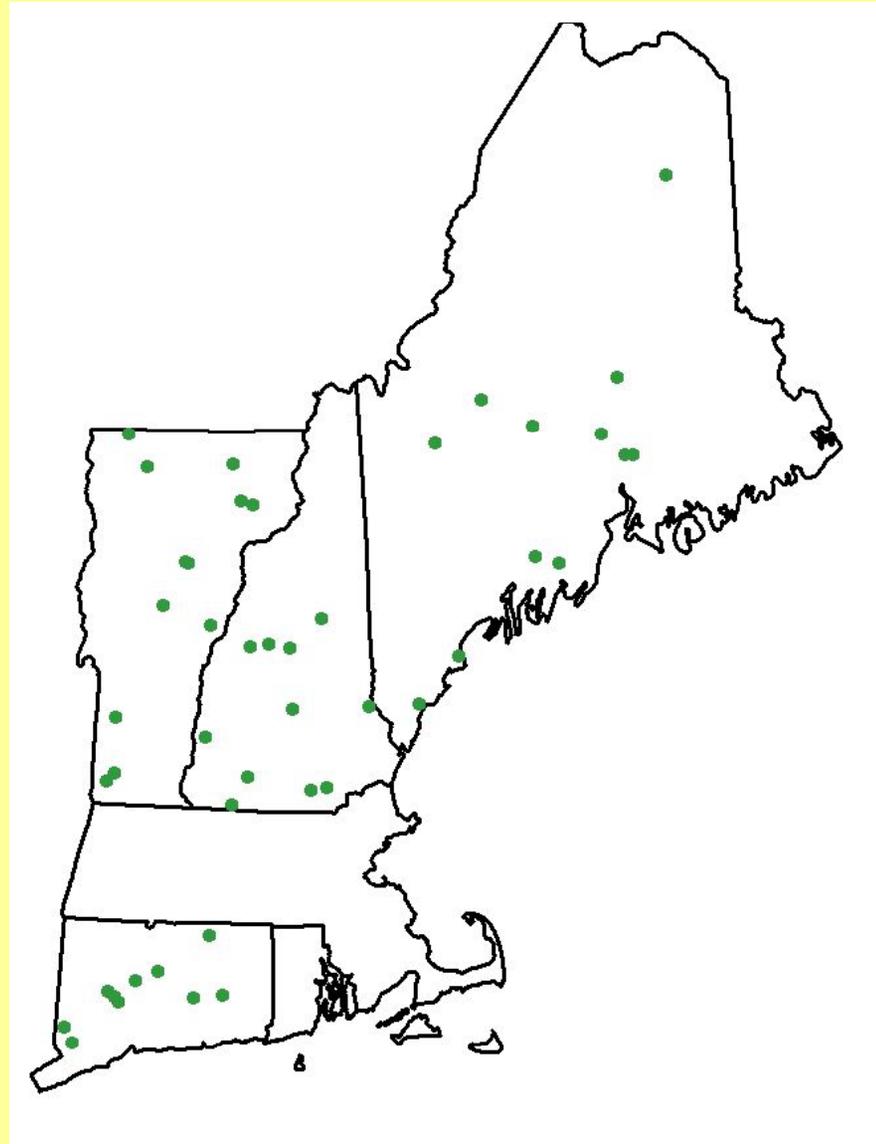
Potential Levels of Comparability:

- 1. Taxonomic Composition:** Do the collection and processing techniques result in a sample with a similar “makeup”? Comparison of raw data. See NEWS report appendix I (Jessup and Gerritsen)
- 2. Ecological Structure and Function:** Do states measure similar ecological components and are these measures comparable? Compare individual components of state indices and apply Biocondition Gradient (BCG) as common translator (See report by Stamp and Gerristen 2009)
- 3. Assessment outcome:** Once samples are processed and ecological attributes computed do states make similar assessment decisions? Check the 303(d) list, but state-specific

Thus, WSA comparability study provided an opportunity to answer the basic question: **Should site “x” be listed as impaired?**

New England WSA sample locations and specifics

- 46 sites sampled in CT, ME, NH, and VT
- Sampling completed according to state protocols
- WSA contractors sampled sites using WSA methods
- Subset of sites also considered “reference” / “impacted” condition



Summary of Methods

CT:

- Rectangular net (800 μ m mesh)
- 12 riffle kicks in fall (Oct.); riffle habitat
- Gridded tray subsampling
- 200 fixed count minimum
- Genus (species) ID endpoint

NH:

- 3 rock baskets
- 6-8 week incubation; 500 μ m sieve; riffle habitat
- Gridded tray subsampling
- 1/4 sample minimum; 100 fixed count minimum
- Genus ID endpoint

VT:

- D-frame net (500 μ m mesh)
- 4 riffle kicks in fall (Oct.); riffle habitat
- Gridded tray subsampling
- 1/4 sample minimum; 300 fixed count minimum
- Genus (species) ID endpoint

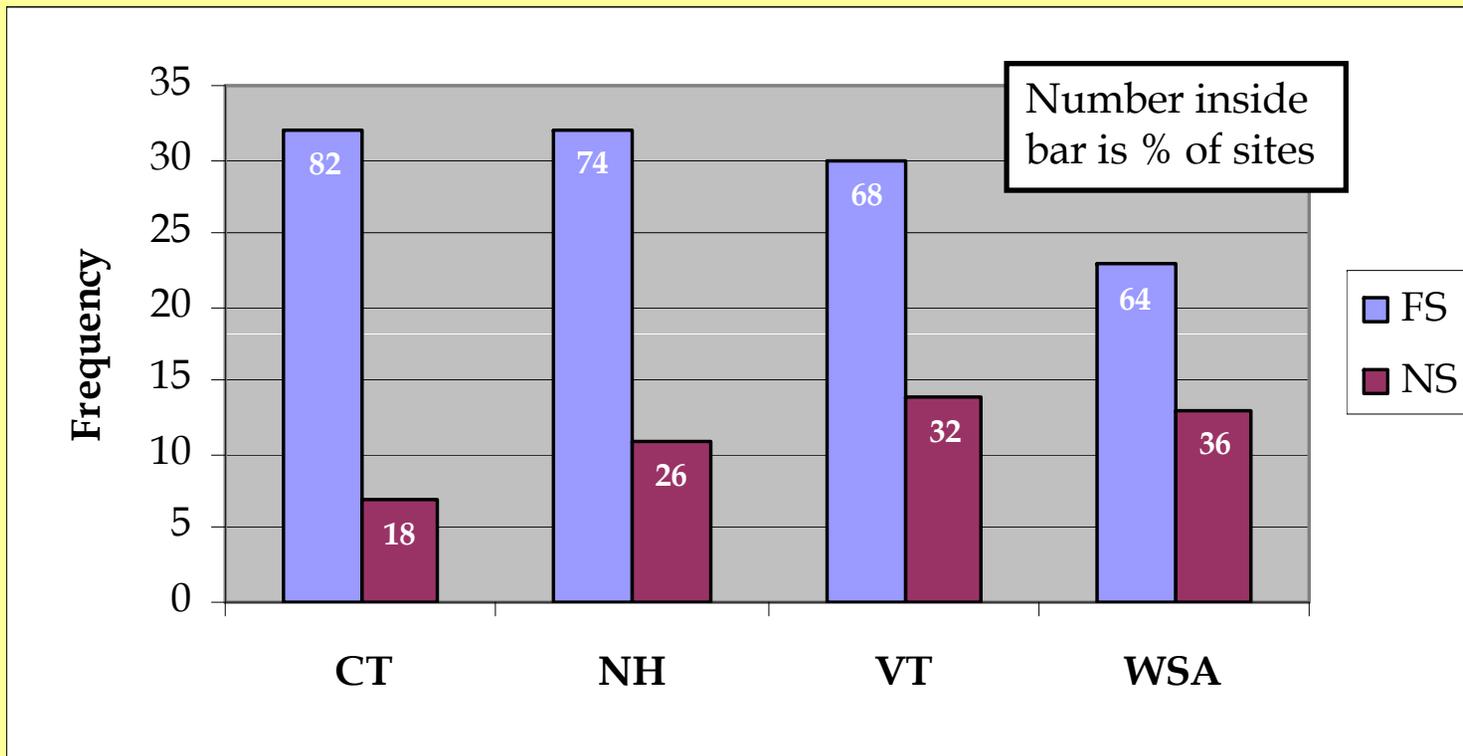
WSA:

- D-frame net (500 μ m mesh)
- 11 kicks stratified along transects; multihabitat
- 500 fixed count
- Gridded sorting
- Genus ID endpoint

Macroinvertebrate Community Condition Evaluation

State	Theshold(s)	Condition "tool"	ALU listing determination
CT	Single w/ grey area requiring BPJ	Multi-metric index; 7 metrics; 0-100; average metric score	Above/Below threshold; threshold based on correspondence with BCG tier rating (Tier 4) + BPJ
NH	Single w/ consideration for reference condition variability	Multi-metric index; 7 metrics; 0-100; average metric score	Above/Below threshold; threshold set at 25 th percent of reference minus 90% confidence interval + BPJ; approximates BCG tier 4
VT	Multiple based on TALU system	Multi-metric evaluation; 8 metrics; evaluated independently	Above/Below "stream class" criteria; Based on evaluation of frequency of metric attainment; Individual metric thresholds based on reference condition; + BPJ
WSA	Single based on narrative rating	Multi-metric index; 6 metrics; 0-100; sum of metric scores	Interpretation; Above/Below "poor" narrative category; Narrative categories established based on reference condition

Overall Comparison of Assessment Outcomes



VT & WSA appear to have lower rate of FS however,

Overall, assessment outcomes do Not differ (chi-square: $p > 0.05$)

Agreement / Assessment Outcome

Does the frequency of disagreement differ between FS / NS outcomes?

Possibilities:

FS-FS - agree

NS-NS - agree

FS-NS - disagree

NS-FS - disagree

Number of Outcomes / Level of agreement:

2 = 2 indices

3 = 3 indices

4 = 4 indices

Complete = all indices agree

Incomplete = 1 or more indices did not agree

Split = equal number of indices / outcome

Total Number of Assessment Outcomes	Frequency
2	4
3	22
4	20

Status	Complete	Incomplete	Split
FS	26	6	---
NS	7	6	---
Total	33	12	1

Agreement more common for FS than NS

Chi-square: $p < 0.10$

MMI score performance (CT, NH, WSA only) (Ode et al. 2008)

Precision

Reference sites	CT	NH	WSA
n	10	9	5
Threshold	45	54-south; 65-north	49-poor
Mean score	72.9	72.2	67.8
CV	0.21	0.23	0.28
Standardized score	1.62	1.11	1.38
CV	0.21	0.46	0.28

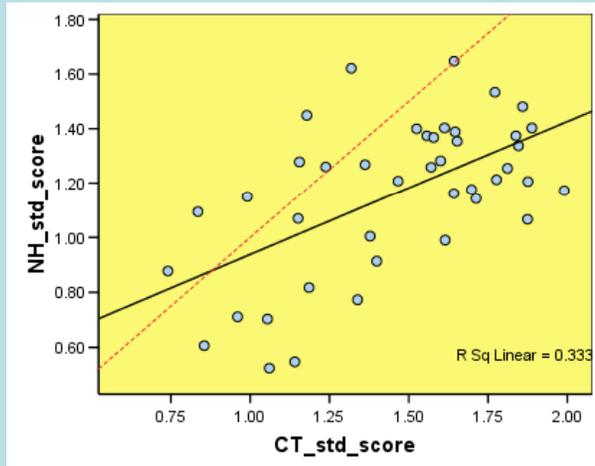
- Few in number
- Actual scores: means, CVs similar
- standardized scores: NH < WSA < CT; NH least precise

Responsiveness - Difference between rescaled MMI score and mean of reference scores

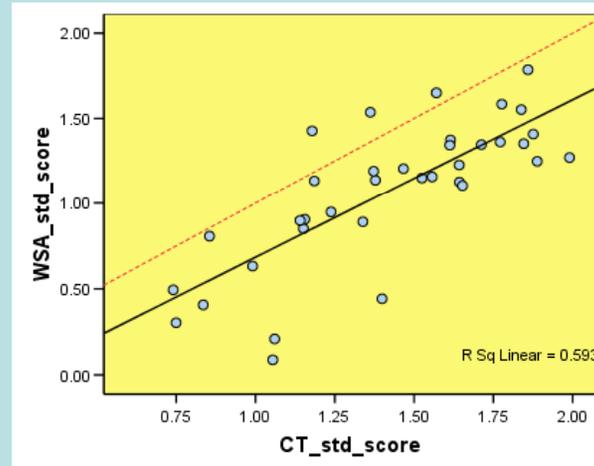
All sites - paired t-test	n	Mean 1	Mean 2	Difference	p
CT vs. NH	42	0.28	0.22	0.06	0.09
CT vs. WSA	36	0.31	0.38	0.07	0.09
NH vs. WSA	34	0.23	0.37	0.14	0.008

- CT \approx NH
- CT \approx WSA
- NH < WSA

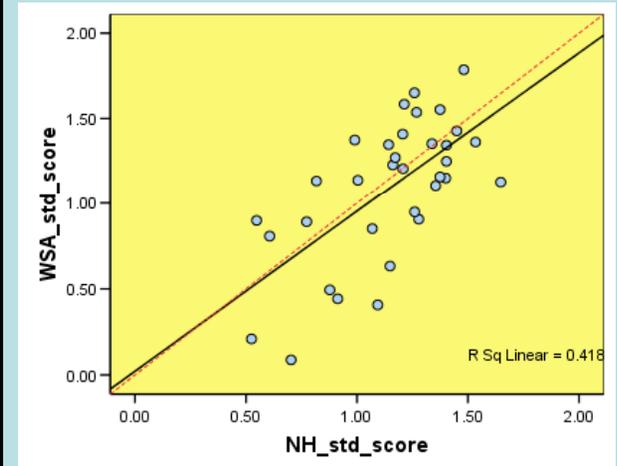
MMI standardized score linear regression



- $R^2 = 0.33$
- $F = 20.0$
- Slope $\neq 1$
- Intercept $\neq 0$
- CT tends to overestimate
- Most variability



- $R^2 = 0.59$
- $F = 49.6$
- Slope = 1
- Intercept = 0
- CT consistently overestimates



- $R^2 = 0.42$
- $F = 23.0$
- Slope = 1
- Intercept = 0
- NH and WSA nearly equivalent according to regression

MMI score conversion using linear regression

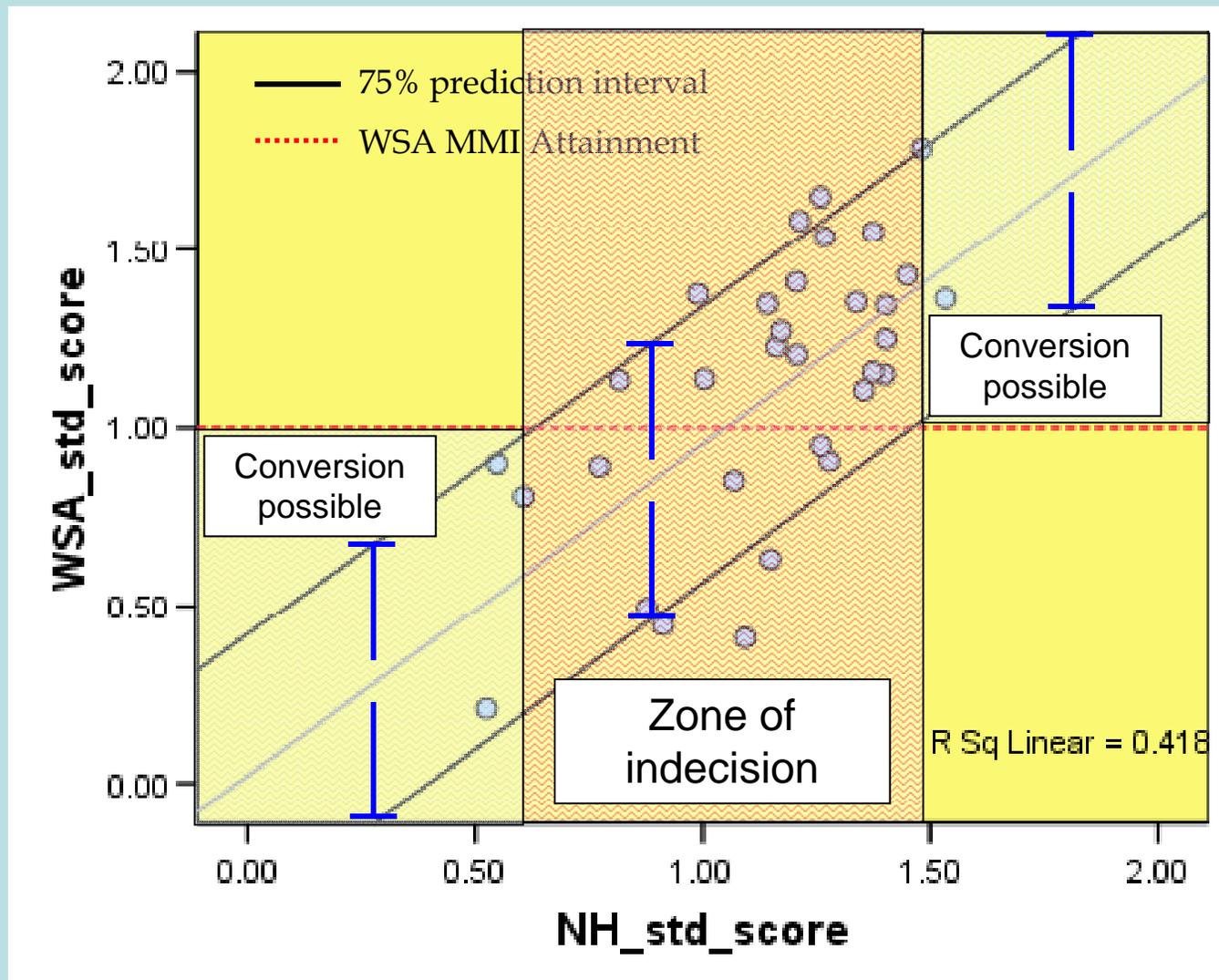
How do predicted assessment outcomes compare to observed outcomes?

Results:

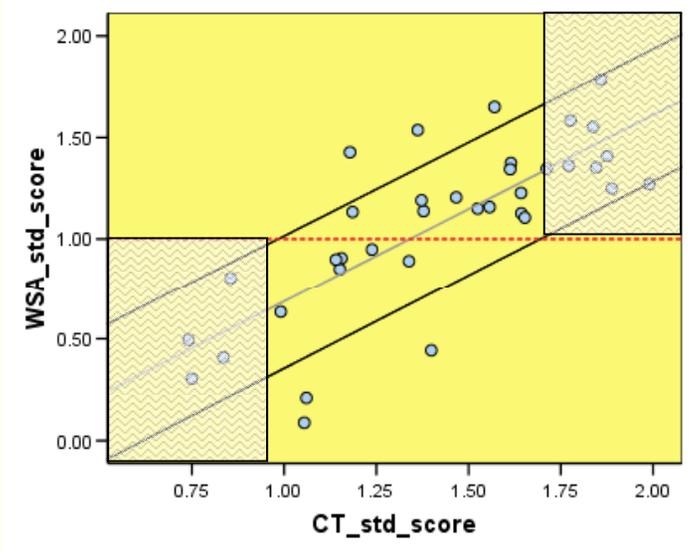
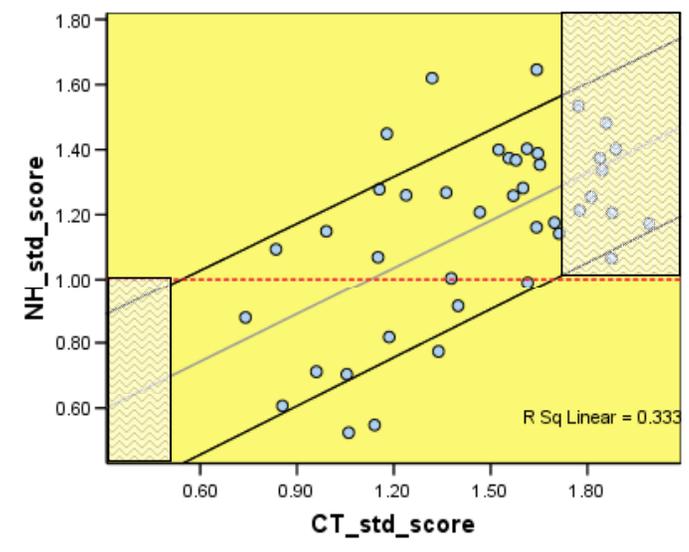
	Agree	Disagree	Act_FS / Pred_NS	Act_NS / Pred_FS
CT / NH				
Pred NH vs. Actual	35 (83%)	7 (17%)	2	5
CT / WSA				
Pred WSA vs. Actual	33 (92%)	3(8%)	2	1
NH / WSA				
Pred WSA vs Actual	26 (77%)	8 (23%)	3	5

Conversion success: CT/WSA > CT/NH > NH/WSA, but does not account for variability

Predicted vs. Actual assessment outcomes taking variability into account



Predicted vs. Actual assessment outcomes taking variability into account



		Actual score needed for predicted assessment (75% PI)		
Actual index score	Predicted index score	FS	NS	Range of unknown outcome
CT	NH	1.70	0.54	1.16
CT	WSA	0.98	1.69	0.71
NH	WSA	1.46	0.63	0.83

Summary of MMI Conversion

1. Use precision and responsiveness of individual MMIs as check on performance
2. Compare MMIs directly using linear regression
3. Estimate strength of relationship using F-value, R^2 , slope and intercept tests
4. Account for variability using prediction intervals and determine range of predicted MMI scores that can be used to make confident assessment outcome calls.

NE example results:

- NH lowest performance (least precise, least responsive)
- CT / WSA best regression (highest R^2 , slope =1, intercept approximates 0), but CT consistently overestimated condition
- CT / NH worst regression (lowest R^2 , slope $\neq 1$, intercept $\neq 0$)
- Zone of indecision smallest for CT / WSA – best opportunity to convert scores into greatest number of assessment outcomes
- Zone of indecision greatest for CT / NH – conversion possible, but wide range where assessment outcomes not possible
- NH / WSA moderate regression result, but lower performance by NH MMI limits conversion into possible assessment WSA assessment calls

Concluding Remarks

- Frequency of FS and NS assessment calls similar among methods, but disagreement was more common for sites in poor condition
- While FS / NS assessment outcomes tend to be relatively consistent, they are coarse end points for reporting condition
- Overall, assessment outcomes indicated VT & WSA were more strict than NH & CT
- MMI score conversion possible but can lead to wide range where assessment outcomes cannot be translated among entities
- MMI score conversion limited to narrative categories that are associated with range of MMI scores (ex. Good, Fair, Poor)
- Comparisons more challenging for non-MMI indices (ME, VT)
- Use of BCG may provide an alternative means for rolling up assessments across entities; requires ability to objectively assign BCG tier (See Stamp and Gerritsen 2009)



Questions, thoughts, suggestions?

David Neils

NH Dept. Env. Services

603/271-8865

david.neils@des.nh.gov