Eastern Kentucky University
Department of Correctional And Juvenile Justice Studies

Center for Criminal Justice Education and Research
Bulletin

CCJER-NIC Bulletin 2
July 2007

National Institute of Corrections Training Academy Evaluation Project, 2005-2006

Participant Evaluation of Trainers

By James B. Wells, Ph.D
Kevin I. Minor, Ph.D.
and J. Stephen Parson, B.A.

This is the second in a series of research bulletins on NIC's Training Academy Evaluation Project (TAEP). Please refer to the February 2007 bulletin subtitled "Participant Demographics, Overall Evaluation of Training, and Applicability Ratings" for a description of both the TAEP and the research bulletin series. That bulletin is available at: www.nicic.org/research.

The current bulletin presents results from analyses of training participants' evaluations of individual trainers. Like the February 2007 bulletin, this bulletin combines data from trainings conducted during 2005 and 2006. As such, the 2005 adult training programs included Correctional Leadership Development (CLD-A) and Management Development

Highlights

- 2,430 completed trainer evaluation surveys were collected from 453 training participants in 20 trainings during 2005-06. Response rates averaged 96.7 percent (Table 1).
- Juvenile training participants provided 70 percent of the completed trainer evaluations, while adult training participants provided about 30 percent (Figure 1c).
- Analyses revealed high levels of satisfaction with 28 of 34 trainers (Figure 8; Table 5). On average, over 98% of evaluations recommended these 28 trainers be used again.
- Six trainers scored lower than the others on a variety of measures (Figure 8; Table 5). Overall about 77% of evaluations recommended these six trainers be used again.
- Five of the six lowest rated trainers gave only one training, while the other 28 trainers averaged about three trainings each (Table 5).
- On average, participants rated trainers in the Juvenile 2006 training programs higher than those in the Juvenile 2005 or Adult 2005-06 trainings (Figures 2 and 4).
- Overall trainers scored an average of 4.43 (SD = 0.55) on an index of the 24 closed-ended items (page 4). Individual trainer averages ranged from 3.47 to 4.78 on this 1 to 5 scale (Table 5).
- Content analysis of open-ended responses produced an open-item index on which trainers scored an average of 2.37 (SD = 2.40) on a –8 to 8 scale. Individual trainers scored from –2.47 to 4.27 (Table 5).
- Trainer evaluations on the closed-item and open-item indices were moderately consistent (see *Two-dimensional Results*, page 8).
- Content analysis of open-ended responses produced very different strength/limitation profiles for above average and below average trainers (Figures 5a and 5b) and identified several dimensions untapped by the closed-ended items (See *Increased Evaluation Breadth*, page 7).
- Overall, the most frequently cited trainer strength, referenced in about 27% of evaluations, was knowledge of their field. The most commonly cited trainer limitation, in over 14% of evaluations, was poor communication or presentation skills (Figure 3).
- About 10% of trainer evaluations cited insufficient time or hurried pace as a trainer weakness (Figure 3), while the closed-ended item addressing this issue was ranked last among the 24 items (Table 2).

CCJER Staff

James B. Wells, Ph.D.
Director and Principal Investigator

Kevin I. Minor, Ph. D.
Co-principal Investigator

J. Stephen Parson, B.A.
Research Associate, NIC Project

Earl P. Angel, B.S.
Research Associate, DJJ Project

Shondra New
Student Research Assistant

Adam Matz
Student Research Assistant

Katherine Flege
Student Research Assistant

Acknowledgements

The National Institute of Corrections Training Academy Evaluation Project is made possible by the support of NIC via Cooperative Agreements 05A28GJF9 and 06PEI01GJM1.

CCJER staff wish to acknowledge the support and cooperation of the many persons who helped make this project possible. Morris Thigpen, Larry Solomon, Chris Innes, Bob Brown, John Eggers, Leslie LeMaster, Launa Kowalcyk and others at NIC have provided essential support for this project. At EKU, Allen Ault, Bruce Wolford, Tommy Norris, and others have been instrumental in creating and nourishing CCJER and other research centers in the College of Justice and Safety, and supporting projects such as the NIC TAEP. We also wish to acknowledge our student support staff, whose daily efforts further the project in so many ways. Finally, we want to express our appreciation to the growing number of NIC Training Academy participants who have taken time out of their busy schedules to graciously share their insights with us.

Although many persons and organizations contributed to the project described in this bulletin, any errors or omissions are those of the authors alone.

The findings and views presented in this bulletin are those of the authors and do not necessarily reflect the positions or policies of the National Institute of Corrections, Eastern Kentucky University, or any other individual or organization.

for the Future (MDF). The 2005 juvenile programs included Training Design and Development (TDD), training for Juvenile Agency Training Directors and Coordinators (JATD), Meeting the Needs of Female Offenders (MNFO), Critical Elements of Re-entry and Continuing Care Systems (CER/CCS), and Juvenile Correctional Leadership Development (CLD-J). Data were collected on all these programs again during 2006, with the exception of TDD and CLD-A, which were dropped from the TAEP, while the New Juvenile Facility Directors (NFD) program was added.

This information is presented in greater detail in Table 1.

As with the February 2007 bulletin, the data for the current bulletin were collected through written surveys. Persons who participated in the trainings mentioned above were asked to complete a Participant's Evaluation of Trainer Survey for each trainer who instructed during the training program. The authors developed the survey instrument while working in close collaboration with NIC Academy staff. Participants were advised that their participation

Table 1: Data Sources and Response Rates, 2005-2006

Panel A: Details by Individual Training

Training		Participants	Trainers	Total Surveys	Returned Surveys	Response Rates
TDD	05-D902	15	4	60	60	100.00%
JATD	05-D801	21	4	84	83	98.81%
MNFO	05-D1001	29	4	116	99	85.34%
MNFO	05-D1002	25	4	100	100	100.00%
CER/CCS	05-D1502	27	4	108	108	100.00%
CER/CCS	05-D1501	24	5	120	120	100.00%
CLD-J	05-D101	27	9	243	242	99.59%
CLD-A	05-M101	22	5	110	110	100.00%
CLD-A	05-M102	25	6	150	150	100.00%
CLD-A	05-M103	22	9	198	191	96.46%
MDF-1	05-R039	34	4	136	136	100.00%
NFD	06-D301	33	4	132	131	99.24%
JATD	06-D801	24	3	72	69	95.83%
MNFO	06-D1001	18	5	90	88	97.78%
MNFO	06-D1002	31	5	155	154	99.35%
MNFO	06-D1003	39	5	195	174	89.23%
CER/CCS	06-D1501	21	5	105	105	100.00%
CLD-J	06-D101	21	8	168	164	97.62%
MDF-2	06-R012	30	3	90	78	86.67%
MDF-3	06-R019	27	3	81	68	83.95%
TOTALS		515[a]	99[b]	2513	2430	96.70%

Panel B: Summaries by Training Type or Year

Training	Participants	Trainers	Total Surveys	Completed Surveys	Response Rates
TDD	15	4	60	60	100.00%
JATD	45	4	156	152	97.44%
MNFO	142	8	656	615	93.75%
CER/CCS	72	6	333	333	100.00%
CLD-J	48	8	411	406	98.78%
CLD-A	69	12	458	451	98.47%
MDF	34	4	307	282	91.86%
NFD	33	4	132	131	99.24%
TOTALS	458[a]	50[b]	2513	2430	96.70%
Juvenile 2005	168	26	831	812	97.71%
Juvenile 2006	187	19	917	885	96.51%
Adult 2005	103	14	594	587	98.82%
Adult 2006	0[c]	4	171	146	85.38%
TOTALS	458[a]	63[b]	2513	2430	96.70%
Adult 05-06	103	14	765	733	95.82%
Juvenile 05-06	351	28	1748	1697	97.08%
TOTALS	454[a]	42[b]	2513	2430	96.70%
2005 All	270	32	1425	1399	98.18%
2006 All	187	22	1088	1031	94.76%
TOTALS	457[a]	54[b]	2513	2430	96.70%

[a] 453 distinct participants (30 MDF participants attended multiple phases of the MDF training; five other participants attended multiple trainings.)

[b] 34 distinct trainers (21 of whom participated in multiple trainings)

[c] Utah MDF involved the same participants in 3 phases of training during 2005-06. Thus all MDF participants are counted in 2005. As the only adult program evaluated in 2006, there were no new participants in adult training programs that year.

Table 2: Participant Evaluation of Trainer, 2005-2006
Closed-ended Survey Items

Survey item asking participant if the trainer:	N	Mean	SD
Clearly defined objectives at the beginning of the training	2419	4.37	0.68
Remained focused on objectives throughout training	2421	4.45	0.64
Accomplished stated objectives	2419	4.39	0.69
Communicated concepts effectively	2419	4.38	0.74
Displayed a thorough knowledge of training material	2421	4.56	0.63
Displayed an appropriate level of preparation	2419	4.51	0.63
Displayed an appropriate level of organization	2418	4.48	0.64
Presented materials at an appropriate pace	2421	4.27	0.80
Effectively utilized training aids	2418	4.33	0.72
Employed a variety of training aids	2417	4.30	0.78
Answered questions clearly	2414	4.38	0.68
Answered questions completely	2414	4.37	0.71
Was enthusiastic about material	2416	4.58	0.63
Demonstrated positive interest in training session	2413	4.61	0.58
Conveyed innovative ideas about topic	2410	4.36	0.73
Conveyed information useful to my job	2411	4.37	0.70
Inspired me to develop new strategies for dealing with topic-related issues	2414	4.34	0.76
Demonstrated an interest in training session	2409	4.56	0.60
Promoted interaction among participants	2414	4.45	0.71
Was able to keep participants involved	2414	4.36	0.78
Was able to keep participants on task	2412	4.34	0.76
Displayed a personal interest in participants and their learning	2412	4.49	0.67
Demonstrated a willingness to assist participants during training	2412	4.52	0.64
Demonstrated a willingness to assist participants outside of training	2408	4.36	0.79

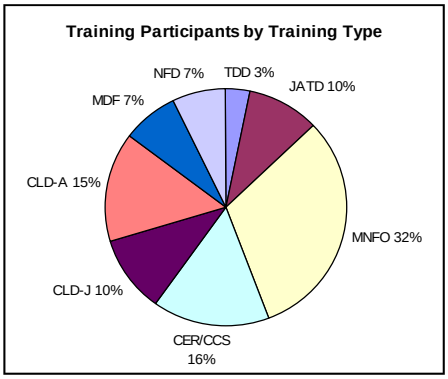

Figure 1a

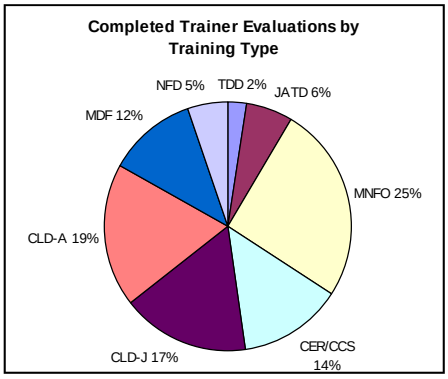

Figure 1b

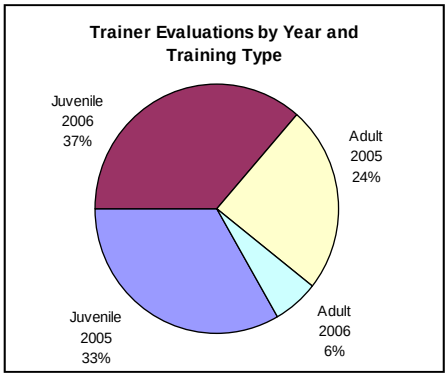

Figure 1c

was entirely voluntary and that they were free to decline to participate, withdraw at anytime, or decline to answer any question that made them uncomfortable. They were also advised that the information they provided would be combined with the information provided by other participants and that their names and other details that might identify them would not be reported. NIC Academy staff or NIC Training Resource Providers administered the instrument on the final day of training, and mailed completed instruments to the Center for Criminal Justice Education and Research (CCJER) for data entry, analysis, and reporting.

One set of closed-ended items on the survey instrument asked training participants to rate trainers on a series of statements regarding each trainer's performance. (See Table 2 on page 2, and *Closed-ended Item Results* below.) The remaining items were open-ended, asking training participants to list the strengths and limitations of each trainer, offer suggestions to the trainers for improvement, and explain why they would or would not recommend the trainer be used again in the same training program. (See *Open-ended Item Results*, page 4.)

As illustrated in Table 1, a total of 2,513 survey instruments were administered to 453 training participants in 20 trainings regarding 34 trainers, and 2,430 completed instruments (96.7%) were collected. Response rates by training ranged from about 84 percent (MDF-3) to 100 percent.

Data sources and response rates are presented for *individual trainings* in Table 1 Panel A. For example, the 27 participants in CER/CCS 05-D1502 each completed evaluations on the four trainers involved in that program, yielding a 100 percent response rate of 108 completed evaluations. Table 1 Panel B summarizes response rates and associated information by *training type*. For example, the 72 total participants in the three CER/CCS trainings completed 333 evaluations of the six trainers involved, for a 100% response rate. Note that direct multiplication is not appropriate for the summarized data in Panel B as not all participants were exposed to all trainers. For example, although six total trainers were used to deliver three CER/CCS trainings, only four or five trainers were used in any one training (see Panel A, Number of Trainers column); thus not all 72 participants evaluated all six trainers.

Figure 1a indicates that almost one-third of all training participants attended MNFO trainings, whereas less than three percent attended TDD. The number of trainers also varied by training, as indicated in column 3 of Table 1, giving rise to a somewhat different distribution of evaluations than participants. For example, while CLD-J participants accounted for about 10 percent of all training participants, (Figure 1a) they completed about 17 percent of all trainer evaluations (Figure 1b). Overall, 1,697 cases (69.8%) came from juvenile trainings, whereas about 30 percent came from adult trainings (Figure 1c). A total of 1,399 cases (57.6%) were from trainings conducted during 2005, while the remainder were from 2006 (Figure 1c).

Findings

Findings from analyses of 2,430 completed trainer evaluation survey instruments are presented in four sections below. Results of participant ratings on the 24 closed-ended survey items are discussed first. Results of content analysis of participant responses to the four open-ended items are discussed next. The third section discusses results from a combined analysis of closed-ended item ratings and open-ended item responses. The final section presents results for individual trainers.

Closed-ended Item Results

Results of participant ratings on the closed-ended survey items are summarized in Table 2[1]. Training participants were asked to indicate on a five point scale the extent to which they agreed or disagreed with each item or statement in regard to each trainer. These ratings were scored as follows: strongly disagree=1; disagree = 2; neutral = 3; agree=4; strongly agree = 5. The number of responses (column N) varies from 2,408 to 2,421 and is slightly less

than the number of completed surveys shown in Table 1 (2,430) because not all participants responded to every item on the instrument. The table provides standard deviations[2] (SD) for each item to specify the average amount of variation within that distribution of scores. The means (averages) in Table 2 reveal that participants' evaluations of trainers were quite favorable overall, falling somewhere between "agree" and "strongly agree" on each of the 24 positively phrased items about the trainers.

Note that the results for each closed-ended item in Table 2 are aggregate or overall measures derived from *all trainer evaluations*. While such measures are helpful in describing general evaluation results, more specific conclusions can be drawn from results for individual trainers, or groups of trainers, such as those involved in a single training (e.g. CLD-J 05-D101) or type of training (e.g. CLD-J, or juvenile trainings generally). To facilitate the analyses necessary to produce such results, mean index scores for closed-ended items were calculated for each trainer evaluation by summing the scores for those items and dividing by 24 (the number of items). The average (mean) of the closed-ended item index scores for trainers in all trainings was 4.43 (N = 2,346, SD = 0.55)[3].

These *closed-item index scores* are used throughout the remainder of the bulletin, alone and in conjunction with other measures, to describe individual trainers and groups of trainers. For example, closed-item index scores for trainers in the adult trainings (mean = 4.38, SD = 0.56) was very close to those for trainers in the juvenile trainings (mean = 4.45, SD = 0.54). The same holds true when comparing trainers in the 2005 trainings (mean = 4.38, SD = 0.56) with those in the 2006 trainings (mean = 4.49, SD = 0.52).

Mean closed-item index scores for smaller groups of trainers, however, reveal somewhat greater variation. For example, as illustrated in Figure 2, the average closed-item index score for trainers in the Juvenile 2006 trainings is 4.55 (SD = 0.48), while the mean index score for trainers in the 2006 adult trainings is 4.10 (SD = 0.62). Likewise, mean closed-item index scores for trainers in the eight training programs listed in Table 3 Panel A range from 4.11 (SD = 0.69) for TDD to 4.66 (SD = 0.44) for NFD. The 20 individual administrations of these training programs shown in Table 4 Panel A reveal mean closed-item index scores for trainers ranging from 3.98 (SD = 0.63) for MDF 06-R012 to 4.75 (SD = 0.37) for MNFO 06-D1001.

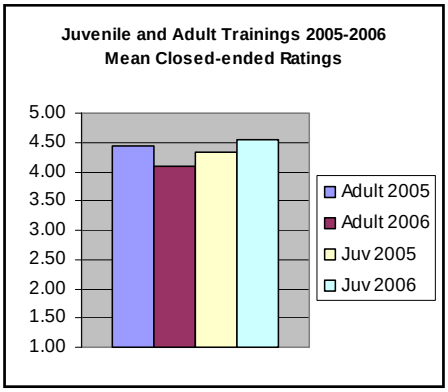

Figure 2

In Tables 3 and 4, trainers or groups of trainers shown in green received above average evaluations overall, while those shown in red received below average evaluations overall. These tables are discussed in further detail in the remainder of the bulletin. Standardized mean index scores (column "Std") in particular, are discussed in the section *Combined Index Results* on page 8.

Open-ended Item Results

Each trainer evaluation survey instrument included four open-ended items (questions or statements) designed to elicit feedback about trainer strengths and limitations[4]. Content analysis[5] of participants' responses to these open-ended items produced rich results that:

1. Identified the trainer strengths and limitations illustrated in Figure 3
2. Provided the basis for calculating an open-item index to facilitate the evaluation and comparison of individual trainers and groups of trainers

Table 3: Average (mean) Trainer Evaluation Scores by Training Type

	Training	N	Panel A: Closed-item Index Mean	SD	Std[1]	Panel B: Open-item Index Mean	SD	Std[2]	Combined Index Std[3]
	NFD	131	4.66	0.44	1.48	4.00	2.01	1.70	1.59
	MNFO	615	4.50	0.56	0.60	2.88	2.55	0.62	0.61
Average	CLD-A	451	4.51	0.52	0.65	2.09	2.20	-0.15	0.25
Average	CER/CCS	333	4.36	0.55	-0.16	2.82	2.36	0.56	0.20
Average	CLD-J	406	4.45	0.47	0.34	2.15	2.11	-0.09	0.12
Average	JATD	152	4.37	0.51	-0.13	2.13	2.59	-0.11	-0.12
	MDF	282	4.17	0.56	-1.23	1.22	1.85	-0.99	-1.11
	TDD	60	4.11	0.69	-1.54	0.65	3.05	-1.54	-1.54

[1] These are standardized mean scores, not mean standardized scores, i.e. they derive from the group means in Panel A (N = 20; mean = 4.391; SD = 0.181), not from the data summarized in Panel A (N = 2,346; Mean = 4.43; SD = 0.55).

[2] These are standardized mean scores, not mean standardized scores, i.e. they derive from the group means in Panel B (N = 20; mean = 2.241; SD = 1.032), not from the data summarized in Panel B (N = 2,430; Mean = 2.37; SD = 2.40).

[3] The standardized combined index scores are simple averages of the standardized closed-item and standardized open-item index scores.

Table 4: Average (mean) Trainer Evaluation Scores by Training

	Training	Code	N	Panel A: Closed-item Index Mean	SD	Std[1]	Panel B: Open-item Index Mean	SD	Std[2]	Combined Index Std[3]
Above Average	MNFO	06D1001	88	4.75	0.37	1.79	4.73	2.20	2.14	1.96
Above Average	NFD	06D301	131	4.66	0.44	1.33	4.00	2.01	1.49	1.41
Above Average	MNFO	06D1003	174	4.54	0.49	0.74	3.89	2.04	1.40	1.07
Above Average	MNFO	06D1002	154	4.64	0.42	1.24	2.60	2.07	0.26	0.75
Above Average	JATD	06D801	69	4.51	0.46	0.61	3.22	2.22	0.80	0.70
Above Average	CER/CCS	06D1501	105	4.37	0.43	-0.12	3.67	1.97	1.20	0.54
Above Average	CER/CCS	05D1501	120	4.50	0.51	0.55	2.88	2.24	0.50	0.53
Average	CLD-A	05M103	191	4.53	0.52	0.68	2.16	2.03	-0.14	0.27
Average	CLD-A	05M102	150	4.53	0.44	0.66	2.11	2.24	-0.18	0.24
Average	CLD-J	05D101	242	4.48	0.42	0.44	2.15	1.96	-0.14	0.15
Average	CLD-A	05M101	110	4.45	0.60	0.27	1.93	2.43	-0.34	-0.03
Average	CLD-J	06D101	164	4.41	0.54	0.07	2.14	2.33	-0.15	-0.04
Average	MDF-3	06R019	68	4.25	0.58	-0.74	2.21	2.12	-0.10	-0.42
Below Average	MNFO	05D1002	100	4.26	0.71	-0.69	1.66	2.66	-0.58	-0.63
Below Average	CER/CCS	05D1502	108	4.19	0.64	-1.02	1.94	2.52	-0.34	-0.68
Below Average	JATD	05D801	83	4.25	0.53	-0.74	1.23	2.54	-0.96	-0.85
Below Average	MNFO	05D1001	99	4.24	0.62	-0.79	1.10	2.33	-1.07	-0.93
Below Average	MDF-1	05R039	136	4.24	0.48	-0.80	0.63	1.36	-1.50	-1.15
Below Average	MDF-2	06R012	78	3.98	0.63	-2.06	1.38	1.96	-0.82	-1.44
Below Average	TDD	05D902	60	4.11	0.69	-1.42	0.65	3.05	-1.47	-1.45

[1] These are standardized mean scores, not mean standardized scores, i.e. they derive from the group means in Panel A (N = 20; mean = 4.394; SD = 0.199), not from the data summarized in Panel A (N = 2,346; Mean = 4.43; SD = 0.55).

[2] These are standardized mean scores, not mean standardized scores, i.e. they derive from the group means in Panel B (N = 20; mean = 2.314; SD = 1.128), not from the data summarized in Panel B (N = 2,430; Mean = 2.37; SD = 2.40).

[3] The standardized combined index scores are simple averages of the standardized closed-item and standardized open-item index scores.

3. Increased the depth of the evaluation by revealing variation between trainers and groups of trainers greater than, yet consistent with, variation revealed by the closed-ended item results
4. Increased the breadth of the evaluation by identifying several dimensions untapped by the closed-ended items
5. In conjunction with closed-item results, provided the basis for extensive instrument revisions[6], resulting in a much shorter, yet more thorough trainer evaluation instrument for future evaluations.

The first four points are discussed in the current section. Additional discussion of the third point, evaluation depth, is provided in *Combined Index Results* and *Individual Trainer Results*, following this section. The final point, instrument revisions, is discussed in *Future Directions* and endnote 6.

Trainer Strengths and Limitations

Content analysis of the open-ended responses revealed the trainer strengths and limitations illustrated in Figure 3. As shown in the figure, the most frequently cited strength, referenced in over 27 percent of all trainer evaluations, pertained to knowledge and skill in the trainer's field of expertise. However, the most frequently cited limitation (in more than 14% of trainer evaluations) related to poor communication and presentation skills. This suggests participants drew a distinction between trainers' knowledge and skill in their field, on the one hand, and the extent to which they successfully communicated these during the training, on the other.

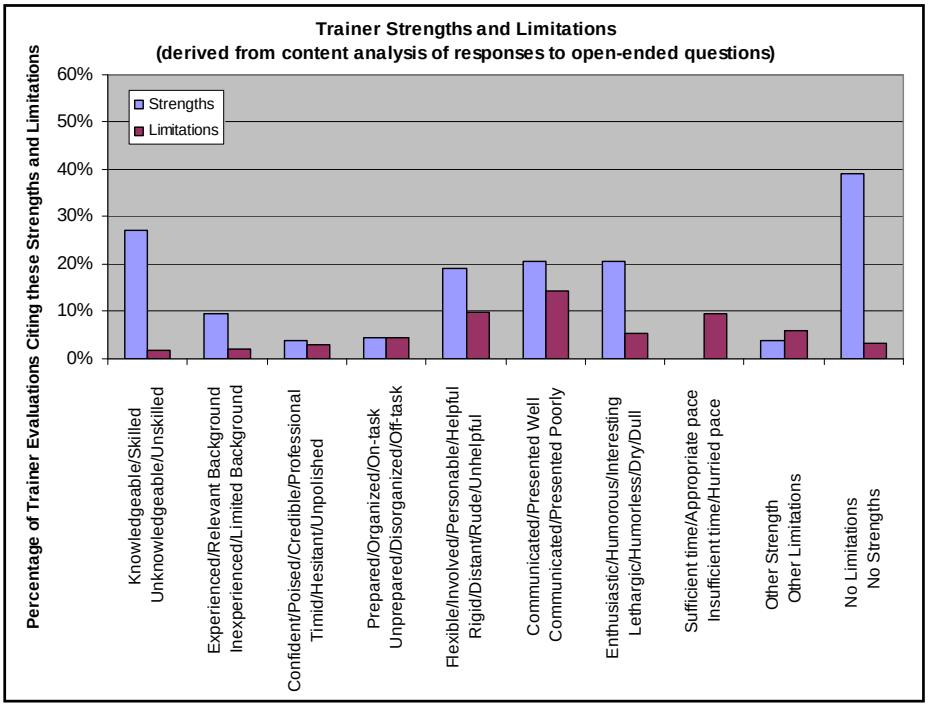

Figure 3

The other most commonly cited trainer limitations included: being rigid, distant, or rude (10.0%); and conducting the training at a hurried pace (9.5%). Other important trainer strengths, each cited in about 20 percent of evaluations, included: communication and presentation skills; enthusiasm and humor; and being flexible and personable. Approximately 39 percent of all trainer evaluations did not identify any limitations, while only three percent identified no strengths.

Open-item Index

While Figure 3 provides an overview of trainer strengths and limitations across all trainers, an *open-ended item index* was formed for each completed evaluation instrument to allow examination of individual trainers or groups of trainers. Open-item index scores provided the basis for many of the other analyses used in the evaluation and are employed throughout the remainder of the bulletin. These index scores were calculated by summing the strengths (each scored 1) and limitations (each scored –1) identified on each instrument during content analysis (see Example A). This method of scoring[7] the four open-ended items on each evaluation produced an open-item index with a range of possible scores from –8 to 8.

The overall mean (average) of the open-item index scores for all trainers was 2.37 (N = 2,430, SD = 2.40).

Example A: Coding and Scoring Open-ended Responses on a Trainer Evaluation

Survey Item:	*What strengths, if any, do you think the trainer brought to this program?*
Response:	Knowledgeable, helpful, down to earth personality
Coded/Scored:	Knowledgeable/Skilled; Flexible/Involved/Personable/Helpful (+2)
Survey Item:	*What limitations, if any, do you think the trainer brought to this program?*
Response:	Difficulty adjusting explanations to accommodate different learning styles, seemed uncertain or unsure of self
Coded/Scored:	Poor Communication/Presentation Skills; Timid/Hesitant/Unpolished (-2)
Survey Item:	*...describe any major reason(s) why you think NIC should consider continuing to use this trainer to deliver the program.*
Response:	Trainer knows the material—is enthusiastic and interesting
Coded/Scored:	Knowledgeable/Skilled; Enthusiastic/Humorous/Interesting (+2)
Survey Item:	*...describe any major reason(s) why you think NIC should consider <u>not</u> continuing to use this trainer to deliver the program.*
Response:	Although knowledgeable in the field, is obviously unskilled as a trainer/presenter
Coded/Scored:	Poor Communication/Presentation Skills (-1)

This trainer scored +1 overall on the open-ended portion of this trainer evaluation.

This indicates that, on average, participants' responses to open-ended items about trainers referenced about two more strengths than limitations. The average open-item index score for trainers in the juvenile trainings was higher than that of the adult trainings: 2.63 (SD =2.47) versus 1.75 (SD = 2.11). Additionally, the mean for trainers in the 2006 trainings was higher than that for 2005 trainings: 3.14 (SD = 2.31) versus 1.80 (SD = 2.31). However, further analysis revealed that higher scores for trainers in the 2006 juvenile training programs accounts for both the differences between program years and between program types (see Figure 4). These trainers averaged 3.36 (SD = 2.27) on the open-item index, while 2005 juvenile program trainers averaged 1.84 (SD = 2.44), 2005 adult program trainers averaged 1.75 (SD = 2.13), and the 2006 adult program trainers averaged 1.77 (SD = 2.07).

Mean open-item index scores for smaller groups of trainers reflect additional variation. For example, average (mean) open-item index scores for the eight training programs shown in Table 3 Panel B ranged from 0.65 (SD = 3.05) for TDD trainers to 4.00 (SD = 2.01) for NFD trainers. Table 4 Panel B indicates that mean open-item index scores for the 20 individual administrations of these training programs ranged from 0.63 (SD = 1.36) for trainers in MDF 05-R039 to 4.73 (SD = 2.20) for trainers in MNFO 06-D1001. Compared to that observed in the closed-item index results (see Figure 2), this is a similar, though more pronounced pattern.

Increased Evaluation Depth

The open-ended items provided training participants greater latitude than closed-ended items in evaluating trainers. Consequently, somewhat greater variation among trainer evaluations is evident in open-item index scores than in the closed-item index scores. For example, as shown in Table 5 (page 8) and discussed in *Individual Trainer Results* (page 9), mean closed-item index scores (Panel A) for individual trainers ranged from 3.47 (SD = 0.63) to 4.78 (SD = 0.31) on a 1 to 5 scale. On the other hand, mean open-item index scores (Panel B) ranged from –2.47 (SD = 2.17) to 4.27 (SD = 2.02) on a –8 to 8 scale.

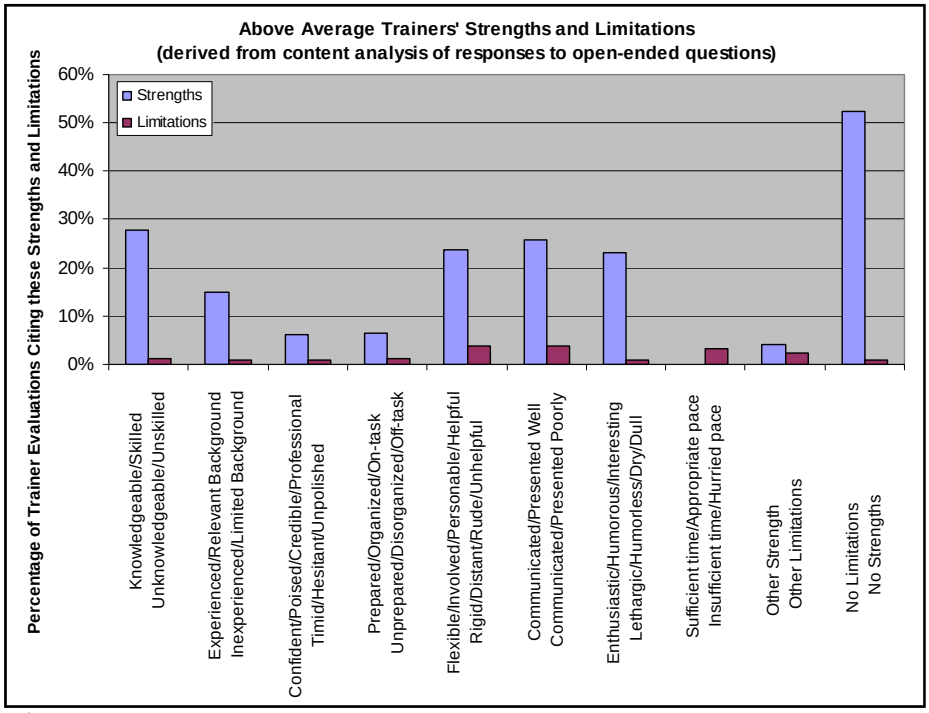

Figure 5a

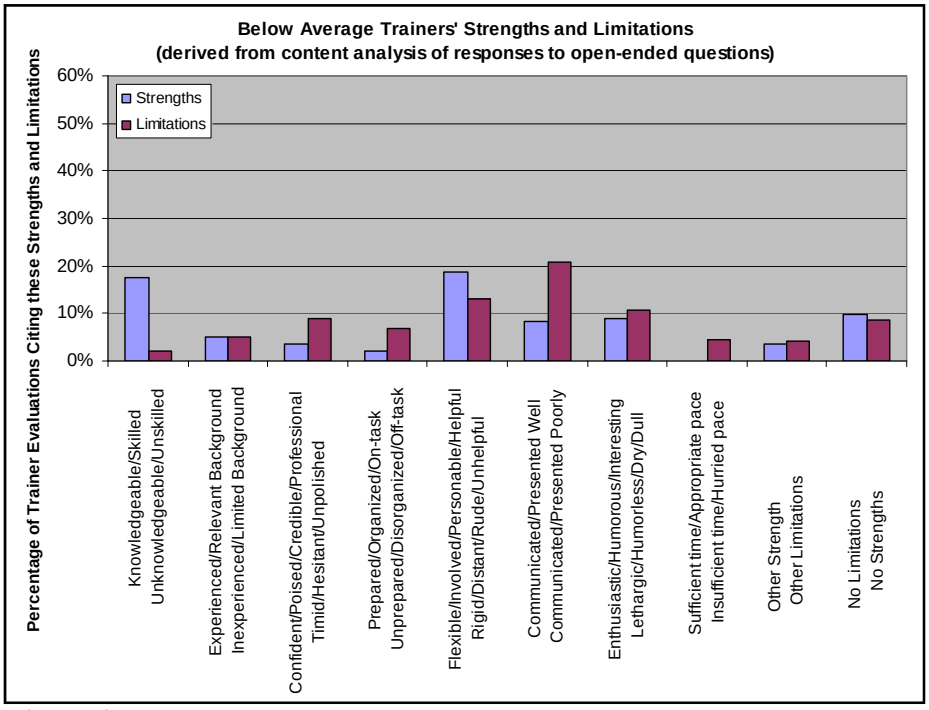

Figure 5b

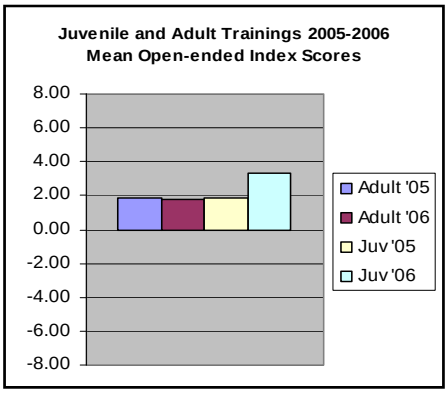

Figure 4

Viewed differently, the mean closed-item index scores varied over 26.2 percent of their possible values, while the mean open-item index scores varied over 39.6 percent of their possible values.

This greater degree of separation between trainers enhanced the depth of the evaluation by providing a means to highlight or magnify the more subtle differences in the closed-item results and also by facilitating examination of trainers on individual dimensions (strengths/limitations). Evaluation depth was also increased in that the open-item index served as a means to cross-check and verify closed-item results.

A variety of statistical analyses were conducted to examine the relationship between the 10 individual dimensions identified through content analysis of open-ended responses (see Figure 3), the open-item index overall, and the closed-item index. Results of all of these procedures were essentially the same; with the exception of confidence/poise, and time/pace, all of the open-ended dimensions were statistically significant.[8] This suggests the other eight dimensions, e.g. knowledge, experience, preparedness, flexibility, communication skills, enthusiasm, etc., were each important in determining training participant evaluation of trainers.

These results, in conjunction with correlation statistics and other findings discussed in *Combined Index Results,* (page 8) suggest a relatively high level of consistency between the two indices. This increased evaluation depth complements, and is complemented by, the increased evaluation breadth discussed next.

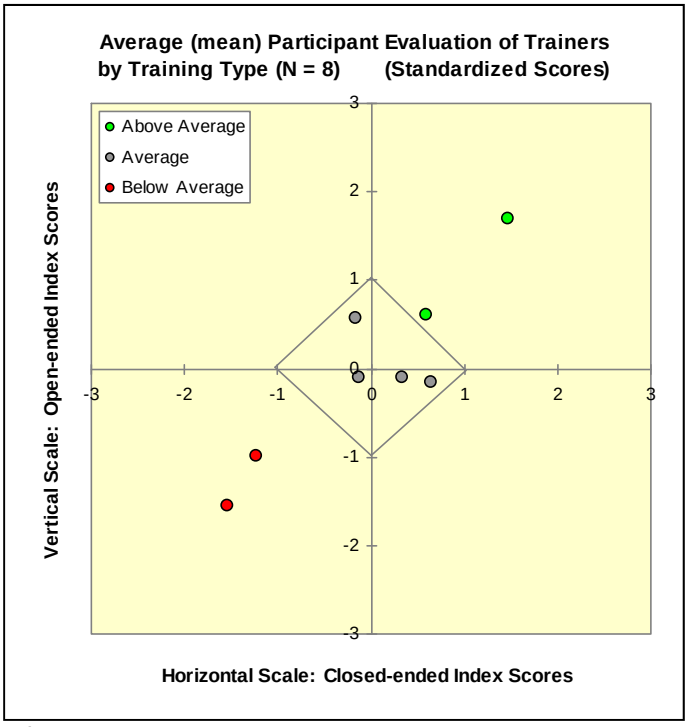

Figure 6a

Increased Evaluation Breadth

Content analysis of open-ended responses revealed two dimensions untapped by the closed-ended items: experience and relevant background (or lack thereof); and confidence/poise/credibility/professionalism (or lack thereof). As illustrated in Figure 5a, and discussed previously, experience and relevant background appears to be an important dimension of trainer evaluation, cited as a strength in about 15 percent of evaluations on above average trainers, while lack of experience was cited as a limitation in less than one percent of these evaluations.

Although the other previously untapped dimension, trainer confidence/poise, was one of two dimensions found to be statistically insignificant, it may nonetheless be of some practical significance. For example, as illustrated in Figure 5a, ***above average trainers were about seven times more likely to receive comments regarding them as confident/poised/credible/professional*** (6.1% of evaluations) rather than timid/hesitant/unpolished (0.9%). Similarly, below average trainers (Figure 5b) were more than twice as likely to receive comments regarding them as timid/hesitant/unpolished (8.8%) versus confident/poised/credible/professional (3.6%).

In fact, most of the statistically significant dimensions show a similar pattern of reversal of the strength-limitation balance between the above average and below average trainers. For example, evaluations of above average trainers were more likely to characterize them as prepared or organized, while evaluations of below average trainers were more likely to refer to them as unprepared or disorganized.

However, exceptions to this pattern were also revealing. For example, ***both above average and below average trainers were far more likely to receive comments regarding them as knowledgeable/skilled (27.9% and 17.4%, respectively), versus unknowledgeable/unskilled (1.1% vs. 2.2%, respectively).*** Even among the lowest rated trainers, being perceived as unknowledgeable or unskilled in their field was the least frequently cited limitation of the 10 identified (Figure 5b). Likewise, both groups of trainers were more likely to receive comments regarding them as flexible/personable, versus rigid/rude. Nonetheless, rigidity/rudeness was the second most commonly cited limitation of below average trainers (13.0% of evaluations) and the most commonly cited though relatively rare limitation of above average trainers (3.9%).

Another notable exception to this pattern is the second dimension found to be statistically insignificant: training time/pace. ***Above average and below average trainers were about equally likely to receive comments suggesting they conducted the training at a hurried pace*** (Figures 5a and 5b). Sufficient training time or appropriate pace was cited as a strength in only about two-tenths of one percent of all trainer evaluations. On the other hand, insufficient time or hurried pace was cited as a limitation in about 10 percent of all trainer evaluations

(Figure 3), and received the lowest overall score of the 24 closed-ended items detailed in Table 2 (see Item 8). ***Thus the dimension of sufficient time/appropriate pace may be of some practical significance for the very reason it was found to be of no statistical significance: it fails to distinguish between trainers, being equally problematic for the highest rated and lowest rated trainers. These findings suggest that factors unrelated to trainer competence, or outside of trainer control, are the source(s) of this limitation.***

Combined Index Results

This section focuses on results that emerge from combined analysis of both indices. When analyzed together, the open-item and closed-item indices of trainer evaluation produce two-dimensional results such as those illustrated in Figures 6a and 6b, and discussed below.

Score Standardization

To facilitate the combined analysis of differently scaled indices, raw mean scores were first standardized. Score standardization converted the closed-item index, scaled 1 to 5, and the open-item index, scaled –8 to 8, to a common scale. Standardizing a group of scores sets the mean score at 0, and the standard deviation[2] at 1, such that scores above the mean become positive numbers, while scores below the mean become negative numbers, regardless of the original scale. For example, standardizing the mean closed-item index scores for the groups of trainers shown in Table 4 Panel A converts the overall mean from 4.394 to 0, and the overall standard deviation from 0.199 to 1 (see Table 4 notes). Consequently, raw scores above 4.394, shown in the "mean" column, become positive numbers in the standardized (Std.) column, while raw scores below 4.394 become negative numbers in the standardized column. Likewise, standardization of the open-item index converted its mean from 2.314 to 0, with raw mean scores above 2.314 becoming positive standardized scores while raw mean scores below 2.314 become negative standardized scores.

Standardization of both indices on the same scale produces scores more appropriate for combination and comparison. For example, standardized open-item and closed-item index scores were averaged to produce overall or combined index scores for trainers in each training listed in Tables 3 and 4 (rightmost column)[9]. These combined standardized index scores determined the final ranking of the groups of trainers listed in those tables.

Two-dimensional Results

Combined analysis of standardized open-item and closed-item index scores produce two-dimensional results such as those illustrated in Figures 6a and 6b. Figure 6a depicts mean trainer evaluation scores on both indices for the eight training programs in the 2005-2006 TAEP. Figure 6b depicts mean trainer evaluation scores on these indices for

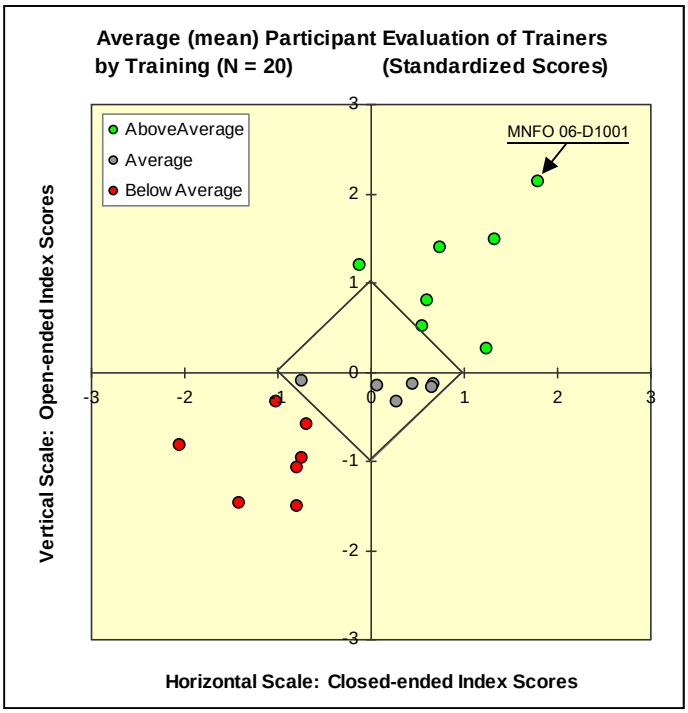

Figure 6b

Table 5: Trainer Evaluations - Mean Raw Scores and Standardized Scores

	Trainer	Number of: Trainings	Number of: Evals	Positive Rec.	Panel A: Closed-item Index Mean	Panel A: Closed-item Index SD	Panel A: Closed-item Index Std.[1]	Panel B: Open-item Index Mean	Panel B: Open-item Index SD	Panel B: Open-item Index Std.[2]	Combined Index Std.[3]
Above Average	6	1	33	100.00%	4.69	0.41	**0.99**	4.27	2.02	**1.41**	**1.20**
	5	1	33	100.00%	4.70	0.50	**1.01**	4.09	2.31	**1.29**	**1.15**
	18	3	68	100.00%	4.78	0.31	**1.25**	3.10	1.96	**0.66**	**0.95**
	28	3	73	100.00%	4.71	0.41	**1.06**	3.22	2.27	**0.73**	**0.89**
	15	4	99	100.00%	4.58	0.45	**0.67**	3.66	2.34	**1.01**	**0.84**
	2	1	15	100.00%	4.66	0.41	**0.90**	3.13	2.00	**0.68**	**0.79**
	13	4	96	100.00%	4.63	0.35	**0.81**	3.26	2.03	**0.76**	**0.79**
	30	1	21	100.00%	4.66	0.44	**0.91**	3.10	1.37	**0.65**	**0.78**
	7	3	80	100.00%	4.51	0.46	**0.46**	3.54	2.24	**0.94**	**0.70**
	21	3	77	100.00%	4.57	0.41	**0.64**	3.06	2.51	**0.63**	**0.64**
	3	3	68	100.00%	4.66	0.34	**0.90**	2.41	1.93	**0.21**	**0.55**
	29	2	44	100.00%	4.46	0.45	**0.31**	3.25	1.81	**0.75**	**0.53**
Average	25	2	45	97.62%	4.38	0.50	**0.06**	3.38	2.06	**0.83**	**0.45**
	34	7	180	100.00%	4.51	0.50	**0.46**	2.75	2.35	**0.43**	**0.45**
	12	8	203	98.28%	4.48	0.55	**0.35**	2.73	2.35	**0.42**	**0.38**
	10	5	114	100.00%	4.58	0.40	**0.65**	2.01	1.93	**-0.05**	**0.30**
	4	5	117	98.25%	4.48	0.54	**0.36**	2.24	2.05	**0.10**	**0.23**
	17	9	233	96.94%	4.41	0.55	**0.14**	2.55	2.44	**0.30**	**0.22**
	27	1	25	100.00%	4.45	0.59	**0.26**	2.28	2.53	**0.13**	**0.19**
	20	3	72	95.65%	4.32	0.67	**-0.12**	2.72	2.42	**0.41**	**0.14**
	31	1	25	95.65%	4.42	0.45	**0.19**	1.96	2.30	**-0.08**	**0.06**
	33	1	21	100.00%	4.41	0.44	**0.14**	2.00	2.00	**-0.05**	**0.04**
	19	3	72	97.06%	4.28	0.54	**-0.24**	2.58	2.56	**0.32**	**0.04**
	26	5	117	100.00%	4.41	0.39	**0.14**	1.30	1.94	**-0.51**	**-0.18**
	16	4	108	97.73%	4.30	0.55	**-0.18**	1.65	2.20	**-0.28**	**-0.23**
	8	2	60	98.04%	4.32	0.45	**-0.12**	1.43	1.76	**-0.42**	**-0.27**
	9	1	27	96.00%	4.18	0.59	**-0.54**	2.04	2.41	**-0.03**	**-0.29**
	14	5	123	97.22%	4.23	0.58	**-0.39**	1.27	1.89	**-0.53**	**-0.46**
Below Average	11	3	83	87.27%	3.92	0.61	**-1.32**	0.81	1.67	**-0.82**	**-1.07**
	23	1	15	93.33%	4.06	0.51	**-0.91**	0.07	2.37	**-1.30**	**-1.10**
	22	1	22	77.27%	3.64	0.67	**-2.17**	-0.45	2.34	**-1.64**	**-1.90**
	1	1	21	68.42%	3.74	0.49	**-1.87**	-1.00	2.05	**-1.99**	**-1.93**
	24	1	25	73.91%	3.65	0.60	**-2.14**	-1.08	1.91	**-2.04**	**-2.09**
	32	1	15	40.00%	3.47	0.63	**-2.66**	-2.47	2.17	**-2.93**	**-2.79**

[1] These are standardized mean scores, not mean standardized scores, i.e. they derive from trainer means in Panel A (N = 34; Mean = 4.360; SD = 0.333), not from data summarized in Panel A (N = 2,346; Mean = 4.43; SD = 0.55).

[2] These are standardized mean scores, not mean standardized scores, i.e. they derive from trainer means in Panel B (N = 34; mean = 2.084; SD = 1.552), not from data summarized in Panel B (N = 2,430; Mean = 2.37; SD = 2.40).

[3] The standardized combined index scores are simple averages of the standardized closed-item and item index scores.

the 20 individual administrations of the eight training programs in the TAEP. In each figure, closed-item index scores are plotted on the horizontal scale, with less favorable or negative scores to the left, and favorable or positive scores to the right. Open-item index scores are plotted on the vertical scale, with negative scores toward the bottom of the graphs and positive scores toward the top.

The closed-item and open-item indices, plotted simultaneously, place the most favorably evaluated trainers to the upper right of the graph, and the least favorably evaluated to the lower left of the graph. For example, the highest rated group of trainers in Figure 6b, those who gave the MNFO training 06-D1001, appears as the green point in the extreme upper right of the figure. This point is located 1.79 standardized units to the right of the center of figure 6b (see Table 4, Panel A, top row), and 2.14 standardized units above the center of the figure (see Table 4, Panel B, top row). Additional details on the results plotted in Figures 6a and 6b are provided in Tables 3 and 4, respectively.

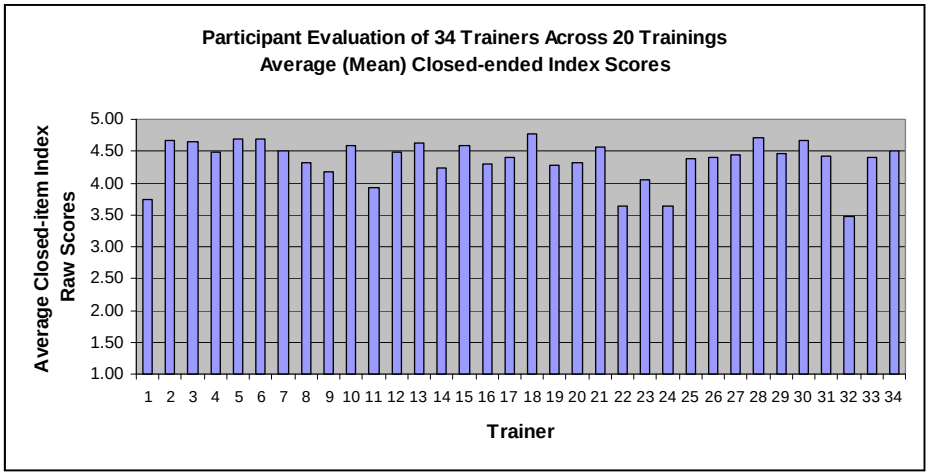

Figure 7a

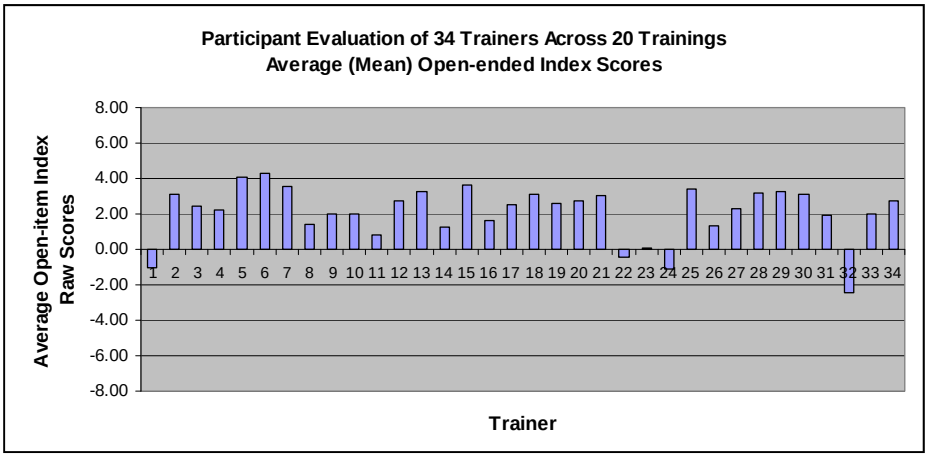

Figure 7b

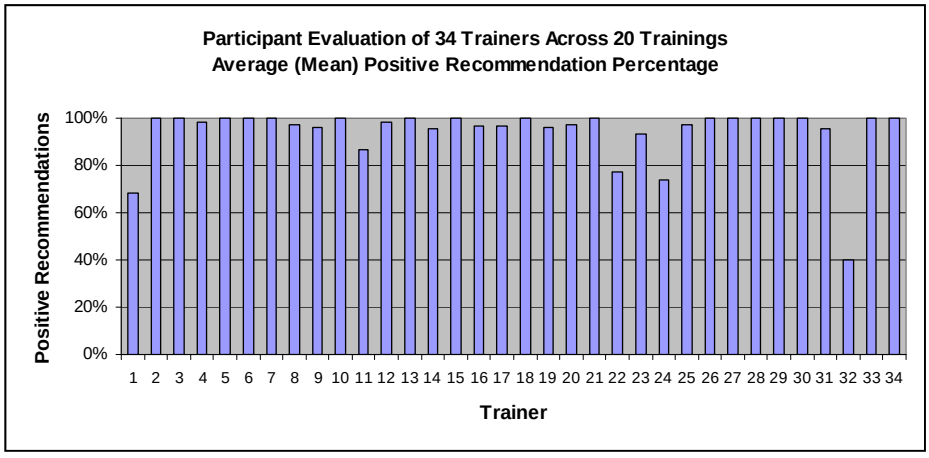

Figure 7c

The somewhat linear pattern of the results in both figures (lower left to upper right) indicates that participants' open-ended responses on trainer evaluations were generally consistent with their ratings on the closed-ended items. Pearson's Correlation, r, a measure of the linear relationship or association between two variables, supports the observation of a moderately strong positive relationship between the closed-item and open-item indices ($r = .536$, $N = 2346$, $p < .001$)[10].

Individual Trainer Results

Thus far the bulletin has focused on trainer evaluation results overall, or results for various groups of trainers. In this section, results are discussed at the level of individual trainers. First, variation *between trainers* overall is examined by comparing the 34 trainers on all available closed-item index results, open-item index results, and positive recommendation levels. Next, variation *between trainers within the same trainings* is discussed. Finally, variation *within trainers* is examined for the 21 trainers who gave multiple trainings by comparing the evaluation results of the same trainer on different trainings.

Variation Between Trainers Overall

A variety of results for each of the 34 trainers are shown in Table 5 (page 8). The most favorably evaluated, or above average trainers, are listed at the top of the table in green shading. The least favorably evaluated, or below average trainers, are listed at the bottom of the table in red shading.

From left to right columns, the table indicates: (a) the trainer code or ID number, (b) the number of trainings in which a trainer participated and the total number of evaluations completed on the trainer, (c) the percentage of respondents who recommended that the trainer should be used again in the same training, (d) raw or un-standardized mean scores on the closed-item index, (e) the standard deviation (average variation) of the closed-item index raw scores, (f) standardized mean closed-item index scores (see *score standardization*, page 8), (g) raw or un-standardized mean scores on the open-item index, (h) the standard deviation (average variation) of the open-item index raw scores, (i) standardized mean open-item index scores, and (j) mean standardized scores for the combined index of closed-items and open-items.

Some variation between trainers is evident in the chart of closed-item index ratings (Figure 7a). While Table 2 (page 2) summarized these results by survey item, averaged across all 34 trainers, Table 5 and Figure 7a summarize these results by individual trainer, averaged across all 24 closed-ended items. As shown in Table 5 Panel A, trainer ratings on the closed-item index ranged from a high (mean raw score) of 4.78, to a low of 3.47 on a 5 point scale.

The open-ended items, as discussed earlier (see *Increased Evaluation Depth*, page 6), provided training participants greater latitude in evaluating trainers. Consequently, greater variation among trainers is evident in Figure 7b, which illustrates trainer scores on the open-item index. For example, 12 trainers were evaluated as having, on average, at least three more strengths than limitations (Table 5 Panel B, raw scores over 3.00), while four others were evaluated as having more limitations than strengths (raw scores less than 0). As shown in Table 5 Panel B, the 34 trainers scored between 4.27 and –2.47 on the open-item index, where possible scores ranged from 8 to –8.

In observing figures 7a and 7b there appears to be a positive correlation between the closed-item and open-item indices, i.e., trainers were generally rated either favorably on both indices, or poorly on both indices. This relationship is most apparent when looking between the figures at the lowest rated trainers on each index, i.e., #1, #22, #24, and #32.

Combined analysis of the closed-item and open-item indices provides a more straightforward illustration of this correlation in Figure 8. In this figure, standardized scores on the closed-item index (horizontal scale) are graphed against standardized scores on open-item index (vertical scale), and each trainer is depicted as an individual data point in two dimensions. To illustrate, consider the evaluation results on trainer #6, the highest rated trainer overall (Table 5, top row). The data point at the upper right of Figure 8, located 0.99 standardized units to the right, and 1.41 standardized units above the center of the figure, graphically illustrates the combined closed-item and open-item evaluation results for this trainer.

The diamond in the center of Figure 8 indicates individual trainers whose combined scores fell into the average range. Green data points to the upper right of the figure represent those trainers with above average combined scores, while red data points to the lower left of the figure represent those trainers with below average combined scores. The linear pattern from lower left to upper right is a graphic representation of the correlation between the closed-item and open-item indices. As reported previously, Pearson's r statistic confirms a moderate positive correlation ($r = .536$, $N = 2346$, $p < .001$)[10] between the two indices[11]. In general, Figure 8 illustrates that 28 of the 34 trainers were evaluated quite favorably overall, while the remaining six were not.

The results depicted in Figure 7c (see also Table 5, Positive Rec. column) were drawn from items asking participants to recommend for or against using each trainer again in the future to deliver the same training. These results can be viewed as a single item or "bottom line" evaluation of each trainer, and a means of cross-checking closed-item and open-item index results[12]. Those trainers shown as above average in Table 5 and Figure 8 achieved 100 percent positive recommendations, while those listed as average achieved about 98 percent positive recommendations. The below average trainers, however, received positive recommendations on about 77 percent of their evaluations.

Variation Between Trainers in the Same Trainings

While trainer evaluations varied between groups of trainers, as discussed earlier in the bulletin, and between individual trainers, as discussed in this section, they also varied *within* trainings. For example, the highest rated of the four trainers involved in TDD 05-D902 scored 3.13

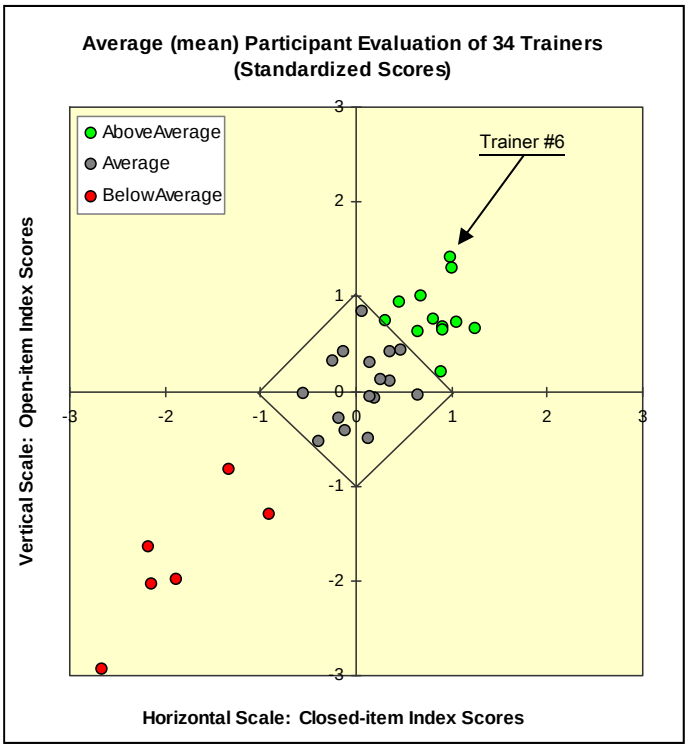

Figure 8

(SD = 2.00) on the open-item index, while the lowest rated trainer scored –2.47 (SD = 2.17). In fact, analysis revealed significant differences[13] between trainer scores on the open-item index in four of the seven below average trainings, and five of the six average trainings shown in Table 4. Likewise, analysis revealed significant differences[13] between trainer scores on the closed-item index in five of the seven below average trainings and five of the six average trainings shown in Table 4.

However, no significant differences[13] between trainer scores on the open-item index were revealed in any of the seven above average trainings. Similarly, analysis revealed no significant differences[13] in closed-item index scores in six of the seven above average trainings. For example, the highest rated trainer of the five involved in MNFO 06-D1001 scored 4.82 (SD = 0.38) on the closed-item index, while the lowest rated trainer scored 4.70 (SD = 0.42). Results[13] reveal that, within the trainings listed as above average in Table 4, participants rated trainers highly, and rather uniformly, on all measures: closed-ended items, open-ended items, and positive recommendations.

These results suggest that, while participants generally distinguished between trainers in a training, they were more likely to do so if they perceived one or more of the trainers to be substandard. Figure 8 provides some support for this observation in that the data points for the 28 trainers evaluated as average or above average are clustered rather closely, while those of the 6 trainers evaluated as below average are considerably more dispersed. Likewise, as illustrated in Figure 5a, above average trainers were rather uniformly evaluated as possessing various strengths and almost no weaknesses, while the evaluations of below average trainers (Figure 5b) were considerably more mixed.

Nonetheless, an unfavorably evaluated trainer can pull down the average trainer evaluation scores for a training enough to shift the training to the below average category. In fact, of the 20 individual trainings, the five lowest rated overall each contained one of the below average trainers. The only training to contain two below average trainers, TDD (05-D902), ranked lowest among the 20 trainings evaluated, despite the fact that the other two trainers were evaluated as above average overall.

Variation Within Trainers (Evaluation of Same Trainer over Several Trainings)

Of the 34 trainers evaluated, 21 took part in multiple trainings. Results indicate that while most trainers' evaluations are fairly consistent from training to training, there are some exceptions. For example, trainers involved in the two 2005 MNFO trainings received below average ratings overall, while trainers in the three 2006 MNFO trainings received above average ratings overall (Table 4). Yet, four of the eight trainers involved in these trainings participated in both years. Three of those four trainers were evaluated as below average overall on the two 2005 MNFO trainings, yet were subsequently evaluated as above average on three 2006 MNFO trainings.

Results suggest that this and similar variation in evaluation scores of the same trainer across trainings stems in part from factors outside of trainer characteristics or performance. Some of these factors may include:

- External constraints imposed by program specialists, curriculum designers, host agencies, etc.
- Situational factors arising from the setting or circumstances of the training, e.g., climate control, flight delays, etc.
- Characteristics of the particular group of participants involved in the training.
- Dynamics arising from the particular group of trainers involved in the training.
- How well a trainer or group of trainers "fit" with a particular group of participants.

Future evaluations and ultimately future trainings may benefit from addressing these and similar research questions. Some suggestions for addressing such questions are discussed in the next section, *Future Directions*.

Future Directions

The trainer evaluation results described in this bulletin can be beneficial in improving future trainings and future training evaluations. These potential improvements, along with the future directions of the evaluation project and the bulletin series are discussed below.

Future trainings may be improved by selecting trainers who possess qualities that extend beyond just expert knowledge of their field. Evaluation results in this bulletin suggest other key dimensions of a successful trainer include: presentation and communication skills; enthusiasm and humor; field experience; confidence and credibility; and being flexible, personable, prepared and organized. Results further suggest that factors beyond trainer characteristics and performance can influence participants' evaluations of trainers. For example, some attention should be given to factors such as: adequate training time to cover the material selected; potential dynamics between the various trainers selected for a training; and how well these trainers are likely to "fit" with a particular group of training participants.[14]

Future trainer evaluations will be improved in several ways based on the evaluation results described in this bulletin.

First, some results suggest that the closed-ended items on the pilot surveys may have influenced the content of the open-ended responses; future research will test for and examine this in more detail. Second, these results provided the basis for extensive instrument revisions, resulting in a survey instrument more thorough than the original, yet with less than half as many items.[6] Finally, results suggest the revised instrument will be more accurate (valid), consistent (reliable), and sensitive (capture more variability) than the pilot instrument.

Future trainer evaluations may also be improved by addressing some deficiencies identified in the evaluation results. For example, although background and demographic data were collected on all training participants, none were collected on the trainers. Thus it was not possible to examine issues such as potential sources of the group dynamics mentioned above, or potential biases in participants' evaluations of certain trainers. Likewise, although data were collected on participants' evaluations of trainers, no data were collected on trainers' ratings of participants' attention, attitude, effort, etc. Thus it was not possible to examine such issues as potential relationships between participants' attitudes or effort levels, and their subsequent ratings of trainers or scores on knowledge tests.

The Training Academy Evaluation Project (TAEP) overall is moving toward a more efficient, streamlined and flexible orientation. For example, in addition to the aforementioned reduced number of items on the trainer evaluation instrument, it has been reformatted to permit the individual evaluation of up to five different trainers on the same instrument. Likewise, most other survey instruments used in the evaluation have been revised and condensed, and in some cases, several instruments have been combined into a single instrument. These and other revisions based on what was learned during the pilot stages of the evaluation project have resulted in a more streamlined evaluation process.

The TAEP is also becoming more flexible and modular as the development of the Evaluation Matrix progresses. The Evaluation Matrix is a grid-like tool and accompanying explanatory text being developed to facilitate the selection of appropriate evaluation type(s) and level of rigor (basic to advanced) for a given training situation. The Evaluation Matrix provides an extensive menu of options from which to choose when arranging for a training program evaluation.

Future bulletins in the series will present additional findings from the various evaluations of the training programs in the TAEP. For example, knowledge gain, behavior change, or organizational change may be examined in depth for a single training program, or across all training programs taking part in the TAEP.

For more information please contact:

Dr. James B. Wells, Director
Center for Criminal Justice Education and Research
Professor, Dept. of Correctional & Juvenile Justice Studies
College of Justice & Safety
Eastern Kentucky University
521 Lancaster Ave, Stratton 105
Richmond, KY 40475

james.wells@eku.edu
(859) 622-1158

Notes

[1] Two versions of the survey instrument were administered over the two year period. Twenty-four of the 29 closed-ended items used were constant across years, and these were the ones analyzed and reported in Table 2.

[2] Generally a small SD indicates most scores fall relatively close to the mean, while a large SD indicates scores are more widely dispersed about the mean. In normally distributed data (approximate bell curve) about 68% of scores will fall within one SD of the mean, while about 95% will fall within two SD of the mean.

[3] The alpha coefficient for the scale was .97. While this indicates a high degree of internal consistency, it also suggests some redundancy across items. As discussed in note 6, factor analysis identified unnecessary items that were then removed.

[4] Two versions of the trainer evaluation survey instrument were administered to training participants during 2005-2006. Both versions included four open-ended items. The earlier version asked:

- What strengths, if any, do you think the trainer brought to this program?
- What limitations, if any, do you think the trainer brought to this program?
- Would you recommend using this trainer again in this program? Please explain.
- What suggestions for improvement would you make to the trainer?

In the later version, the first two of these open-ended items were included as is, but the final two items were modified to read:

- Based on the quality of your learning experience in this program, describe any major reason(s) why you think NIC should consider continuing to use this trainer to deliver the program.
- Based on the quality of your learning experience in this program, describe any major reason(s) why you think NIC

should consider not continuing to use this trainer to deliver the program.

[5] Content analysis of participants' responses to the open-ended items consisted of:

- A cursory review of the data by a research assistant who developed a set of draft categories, e.g. "knowledge, experience, preparedness, … " based on patterns observed in the data
- The selection of a 10 percent random sample of the data that was then coded or collapsed (assigned to draft categories) by three independent research assistants who achieved a 69 percent inter-rater agreement (raters were compared against each other in pairs)
- The draft categories were refined based on problem areas identified in the initial coding run
- Research assistants then closely reviewed and coded or collapsed all data into these final (revised) categories
- A second 10 percent random sample of the data was then coded by three independent research assistants who achieved an 80 percent inter-rater agreement

[6] Factor analysis identified 16-18 redundant closed-ended items on various versions of the instrument. These were removed. Content analysis of open-ended responses revealed several dimensions untapped by closed-ended items, and some redundancy among open-ended items. Closed-ended items were added to address the untapped dimensions; open-ended items were reduced from four to one. Overall the trainer evaluation instrument was reduced from about 32 items to about 14. Taken together with reformatting the instrument and other revisions, future training participants will complete a single one page survey to evaluate up to five trainers. Prior to this round of instrument revisions, five two-page pilot surveys would have been required to evaluate those same five trainers.

[7] The content analysis procedure was designed to identify up to three strengths within responses to each of the two "strength eliciting" items (6 maximum) on each instrument. Although these items were worded to elicit strengths, responses such as "this trainer had no strengths…" were coded as limitations. Thus possible scores on each of the two "strength items" ranged from –1 to 3, (–2 to 6 for both items). Likewise, the procedure identified up to three limitations in each of the two "limitation eliciting" open-ended items (–6 minimum) on each instrument. Although these items were worded to elicit limitations, responses such as "this trainer had no limitations…" were coded as strengths. Thus possible scores on each of the two "limitation items" ranged from –3 to 1, (–6 to 2 for both items). This method of scoring the four open-ended items on each evaluation produced an open-item index with a range of possible scores from –8 to 8.

[8] Given the highly skewed nature of most of the data, logistic rather than multiple regression was utilized to determine which of the ten variables were a significant predictor of overall trainer evaluation score, while controlling for the effects of the other nine variables. (A multiple regression model with skewed data was conducted nonetheless; almost identical results were achieved with this model: $R^2 = .29$). In order to do this, the overall trainer evaluation data were collapsed into two groups, one above the median score of 4.54 (High Trainer Evaluation Score Group) and one at the median score and below (Low Trainer Evaluation Score Group). Data screening led to the elimination of several outliers.

Regression results indicated the overall model fit of eight predictors (confidence/poise and appropriate pace/time did not enter the model) was questionable (-2 Log Likelihood = 2465.224) but was statistically reliable in distinguishing between the two groups (X^2 = 449.712, df = 8, $p < .0001$). The model correctly classified 68.5% of the cases. Wald statistics indicated that all eight variables significantly predicted group membership. All odds ratios were above 1.0, indicating that as each of the eight variables increased by 1, the odds of being classified in the high evaluation score area increased by the respective ratio (two variables, Preparedness and Communication, had odds ratios above 2.0). R_L^2 was .26 and is analogous to the multiple linear regression R^2 indicating amount of variation explained. (A logistic regression model with outliers not removed was also conducted; almost identical results were achieved with this model: $R^2 = .27$).

[9] Standardized scores are also employed in the presentation of combined index results in Figure 8, and Table 5.

[10] Pearson's Product Moment Correlation, r, is a measure of the strength and direction of the linear relationship between two variables. Values of this statistic range from –1 to 1. Negative r values indicate a negative or inverse relationship, i.e., as one variable increases, the other decreases. Positive r values indicate a positive or direct relationship, i.e., both variables increase or decrease together. r values near –1 or 1 indicate very strong relationships while values near 0 indicate very weak relationships. The p value listed with r, in this case $p < .001$ (less than 1 in 1,000), indicates the probability r reflects a random or chance occurrence rather than a true relationship or correlation. Note that Pearson's Correlation, r, is an appropriate measure of a linear relationship between two normally distributed variables measured at the interval or ratio level. However, in the social sciences, it is customary to apply Pearson's r also to non-normal data, as long as there are no serious outliers or departures from linearity. Likewise, social scientists frequently estimate or approximate interval/ratio level of measure by creating scales or indices from data measured at the ordinal level, as was done in this bulletin with the closed-item and open-item indices.

[11] While this correlation indicates some consistency between closed-item ratings and open-item comments, it does not appear to suggest excessive duplication. Rather, taken together with results presented thus far, it suggests the two indices address somewhat different dimensions of trainer evaluation.

[12] Due to item revisions between 2005 and 2006, a small portion of the open-item index and recommendation index overlap.

[13] One-way ANOVAs were used to test for significant differences between evaluations of trainers within the same training. Levene's test was used to verify homogeneity of variances. Tukey HSD was the post hoc used to identify trainers with significantly different evaluation scores when the ANOVA was significant. Forty sets of these analyses were run; one for the closed-item index scores and one for the open-item index, for each of the 20 trainings. Results are far too lengthy to be reproduced in this bulletin, but are available from the lead author.

[14] Considerable insight into group dynamics such as those mentioned here can be found in the academic discipline of social psychology.