Motivational Interviewing training for child and family social workers in Finland:

An exploratory evaluation study

Supplementary material

1. Descriptive results

1.1. Descriptives of practitioners in the data set (n=22) (Table S1)

	n	%
Work unit		
Assessment	3	13,6
Family social work / child in need	5	22,7
Child protection, in-home services	5	22,7
Child protection, out-of-home care	2	9,1
Child protection, combined	7	31,8
Work experience in social work		
1-2 years	4	18,2
3-5 years	6	27,3
6 years or more	12	54,5
Work experience in child and family social work		
1-2 years	5	22,7
3-5 years	7	31,8
6 years or more	10	45,5
Degree		
Licensed social worker	20	90,9
Substitute qualification	2	9,1
Previous MI training		
None	11	50,0
Introductory course or independent learning	9	40,9

1–2 day training or longer	2	9,1

1.2. Key issues and concerns (Table S2)

Key issues and concerns	N
School attendance	12
Child custody and meeting arrangements	9
Child's behaviour	7
Parent's coping	6
Violence, abuse	5
Parenthood, raising children	5
Substance abuse (parent)	5
Health (child)	4
Economical situation, benefits	4
Relationship between divorced parents	4

1.3. The number of meetings (Table S3)

	T1			T2		
	n	M	SD	n	M	SD
Scheduled Meetings	346	1.38	1.13	221	1.36	1.06
Canceled Meetings	58	0.23	0.48	25	0.16	0.39

Completed Meetings	286	1.14	1.00	184	1.14	1.00

1.4. Meeting locations (Table S4)

	T1		T2	
	n	%	n	%
Office	101	34.5	75	39,9
Home	58	19.8	43	22,9
Remote	73	24.9	30	16,0
Other	48	16.4	36	19,1
Total	**280**	**95.6**	**184**	**97,9**
Missing	**13**	**4.4**	**4**	**2,1**
Overall Total	**293**	**100.0**	**188**	**100,0**

1.5. Frequency and Percentage of Recorded Meetings at T1 and T2 (Table S5)

	T1		T2	
Recorded	n	%	n	%
No	213	73,2	154	81,9
Yes	78	26,8	34	18,1

1.6. Reasons for not recording (Table S6)

	T1		T2		T1+T2	
	n	%	n	%	n	%
Consent requested, all declined	19	6.5	11	5.9	30	8,5
Consent requested, some declined	19	6.5	12	6.4	31	8,7
No consent requested, more than 5 participants present	55	18.8	38	20.2	93	26,2
No consent requested, language other than Finnish, Swedish, English, or Russian	20	6.8	26	13.8	46	13,0
No consent requested, child under 12 present without a parent/guardian	7	2.4	3	1.6	10	2,8
No consent requested, anticipated harm to client	27	9.2	29	15.4	56	15,8
Other reason	55	18.8	34	18.1	89	25,1
Total	**202**	**68.9**	**153**	**81.4**		
Missing	**91**	**31.1**	**35**	**18.6**		
Overall Total	**293**	**100.0**	**188**	**100.0**		

1.7. Number of participants in meetings (Table S7)

T1		T2	
M	SD	M	SD
3,55	2,21	4,04	2,20

1.8. Number of participants in meetings, excluding the social worker (Table S8)

	T1		T2	
	n	%	n	%
1	34	11,7	17	9,0
2	45	15,5	25	13,3
3	65	22,3	44	23,4
4	39	13,4	28	14,9
5 or more	90	30,9	71	37,8
Missing	18	6,2	3	1,6

1.9. Meeting participants (Table S9)

	T1		T2	
	n	%	n	%
Child only	9	3,1	8	4,3
Custodian only	22	7,6	10	5,3
Child and custodian only	14	4,8	6	3,2
Child and custodian or other parent or	147	50,5	115	61,2

other close

person only

2. Additional analyses

2.1. Paired samples t-test

2.1.1. Statistical power

The attained statistical power can be observed in Figure S1. The curve displays statistical power of the paired t-tests, given an unknown actual effect size. Given the point estimate meta-analytic effect size of Schwalbe & Oh (2014), the probability of a single t-test observing a statistically significant effect at $p < 0.05$ (not adjusted for multiple comparison), is 62% for the real service users and 45% for the simulated interviews.

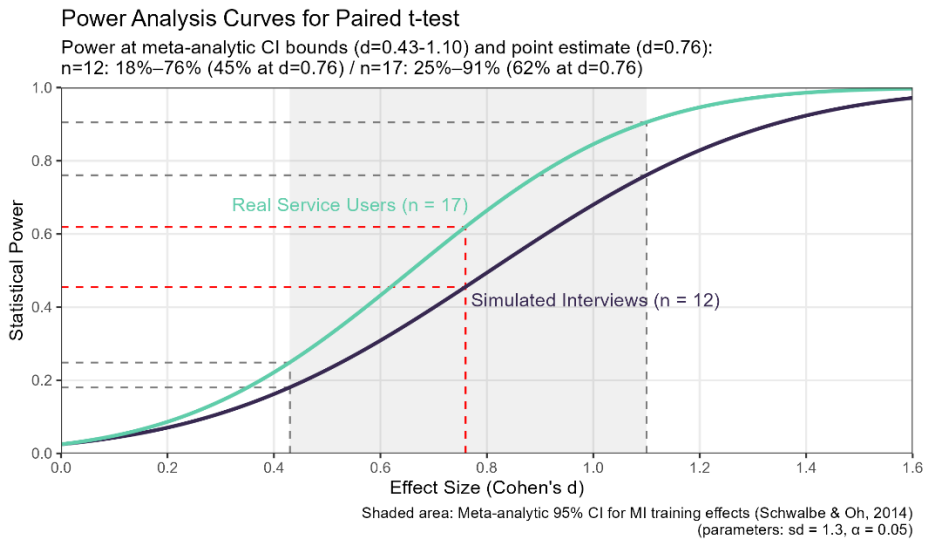

Figure S1. Power analysis curves for a paired t-test.

2.1.2. Assumptions

The difference scores were normally distributed when assessed by Shapiro-Wilk's test for evocation (p = .314), collaboration (p =.063), purposefulness (p = .141) and child focus (p = .074), but not for autonomy (p < .001), empathy (p = .032) or clarity of concerns (p = .033). In addition, the normality of the variables remained unclear in the visual inspection.

There were no outliers in the data for evocation, purposefulness, child focus, autonomy, empathy or clarity of concerns as assessed by inspection of a boxplot. One outlier was detected for collaboration and three for child focus. An examination of their values did not indicate any extremes, and they were retained in the analysis.

2.1.3. Sensitivity analyses / robustness checks

2.1.3.1. Paired Samples t-test simulated meetings (Table S10)

	T1		T3		*t*(df)	*p*
	M	*SD*	*M*	*SD*		
Evocation	2.89	.601	2.89	.782	.000(8)	1.00
Collaboration	3.50	1.00	3.83	.937	1,773(11)	.104
Autonomy	3.50	.522	3.42	.515	-.364(11)	.723
Empathy	3.00	.853	3.75	.965	2.462(11)	.032
Purposefulness	3.25	.452	3.50	.798	1.149(11)	.275
Clarity of concerns	3.75	.622	3.25	.866	-1.915(11)	.082
Child focus	3.33	.492	3.25	.622	-.321(11)	.754

* $p < .05$.

2.1.3.2. Paired Samples t-test simulated and real meetings combined (Table S11)

	T1		T3		*t*(df)	*p*
	M	*SD*	*M*	*SD*		
Evocation	3.00	.679	2.79	1.051	-.641(13)	.533
Collaboration	3.38	.942	3.66	.897	1,440(28)	.161
Autonomy	3.28	.455	3.38	.494	.828(28)	.415
Empathy	2.93	.799	3.62	.820	3.576(28)	.001
Purposefulness	3.34	.614	3.48	.911	.701(28)	.489
Clarity of concerns	3.45	.827	3.41	1.086	-.197(28)	.846
Child focus	3.21	.509	3.13	.850	-.440(23)	.664

* $p < .05$.

2.1.3.3. Wilcoxon Signed-Rank Tests (Table S12)

	T1		T3		Z	*p*
	M	*SD*	*M*	*SD*		
Evocation	3.20	0.84	2.60	1.52	-.707	.480
Collaboration	3.29	0.92	3.53	0.87	-.988	.323
Autonomy	3.12	0.33	3.35	0.49	-1.633	.102
Empathy	2.88	0.78	3.53	0.72	-2.221	.026
Purposefulness	3.41	0.71	3.47	1.01	-.372	.710
Clarity of concerns	3.24	0.90	3.53	1.23	-1.387	.166
Child focus	3.08	0.52	3.00	1.04	-.276	.782

* $p < .05$.

2.1.4. Benjamini-Hochberg Procedure (Table S13)

Benjamini-Hochberg Procedure calculated with a false discovery rate of 5%, 10% and 20%.

$(i/m) \times Q$

i = the individual p-value's rank,
m = total number of tests,
Q = the false discovery rate

Variable	p-value	Rank	5% FDR Critical Value	10% FDR Critical Value	20% FDR Critical Value
Empathy	.023	1	.0071	.0143	.0286
Autonomy	.104	2	.0143	.0286	.0571
Clarity of concerns	.172	3	.0214	.0429	.0857
Collaboration	.450	4	.0286	.0571	.1143
Evocation	.468	5	.0357	.0714	.1429
Child focus	.777	6	.0429	.0857	.1714
Purposefulness	.848	7	.0500	.1000	.2000

3. Exploratory Data Analysis (EDA)

3.1. Person-level skills

Figure S2 below depicts the raw scores for the SWIM-coded skills, providing the necessary context to interpret the aggregate-level results presented in the main manuscript and tables S10-S12.

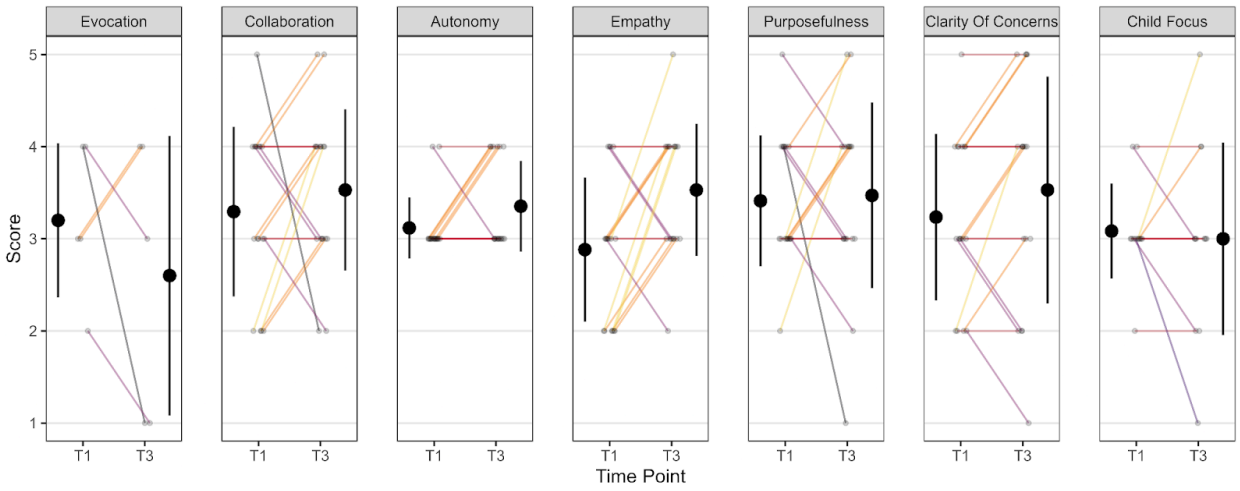

Figure S2. SWIM-assessed skill values. Black points with error bars on the *sides* of each panel represent means and standard deviations for the skill category. Transparent points connected by lines in the *middle area* of each panel, represent values for each individual participant. Points are jittered on the horizontal axis to avoid overplotting. Brighter line colour indicates more improvement between the two time points.

Figures S3-S4 below illustrate the person-level analysis conducted. In step 1 of the exploratory data analysis, such plots were examined to qualitatively group the individual participants into different performer categories.

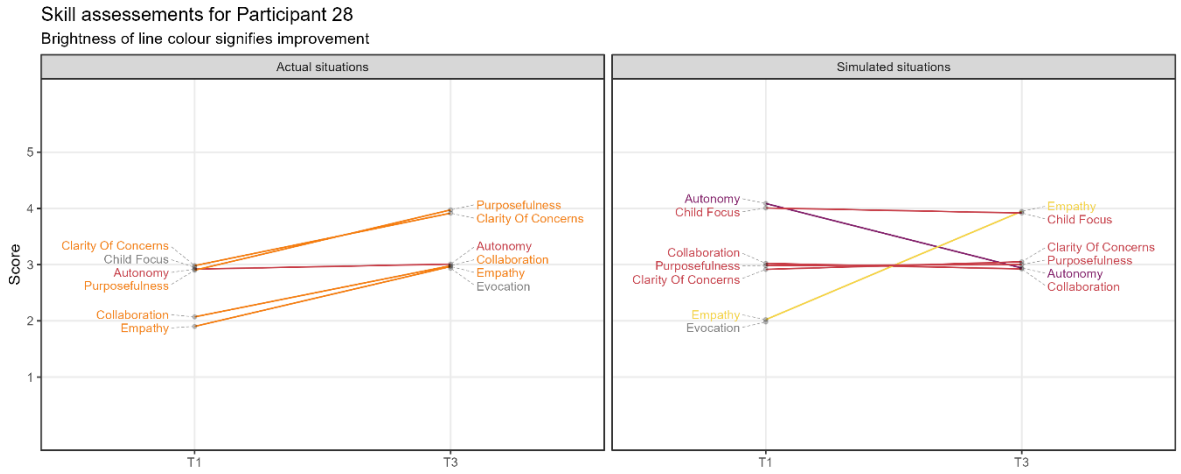

Figure S3. A participant categorised as an “improver” in step 1 of exploratory data analysis.

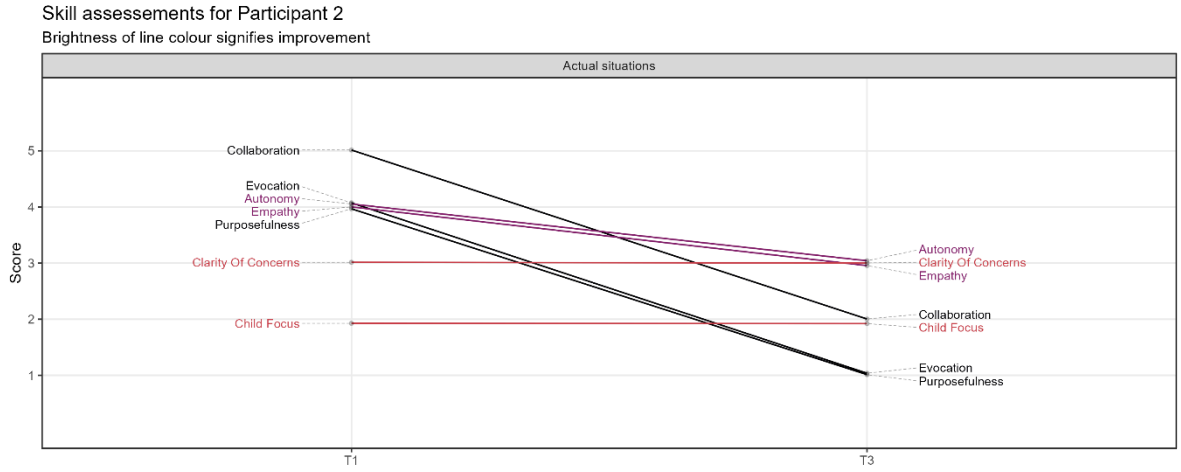

Figure S4. A participant categorised as an “decliner” in step 1 of exploratory data analysis.

3.2. The relationship between SWIM scores and survey responses

3.2.1. Visual analysis

Figure S5 below depicts the relationship between mean baseline SWIM scores and changes between the two time points. We chose a 0.5-point change threshold to indicate improvement, but as there was no theoretical rationale behind the sharp 0.5-point threshold, participant ID 28 was considered as an improver in the follow-up analysis. This is a conservative decision, as it weakens our results somewhat; omitting ID 28 from the improver category leads to 81% of

variables depicting the hypothesised sequential order (i.e. decliners being the lowest category, other being in the middle, and improvers in the top for each variable). As reported in the manuscript, including ID 28 leads to 79% of variables displaying the sequence.

Three distinct groups were hence derived:

- Improvers (IDs 11, 12, 17, 19, 26, 28, 29, 32, 35)
- Decliners (IDs 2, 8, 10, 14)
- Others (IDs 1, 3, 4, 6, 13, 15, 16, 22, 25)

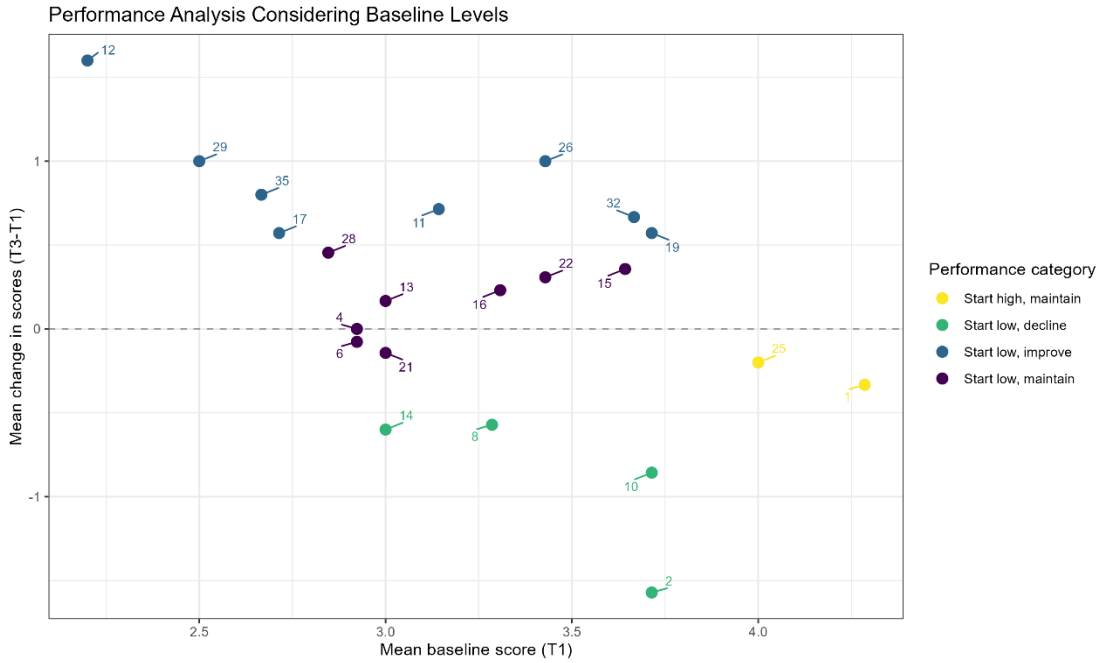

Figure S5. Relationship between mean baseline scores (T1) and mean change in scores (T3-T1) across MI skills. Points represent individual participants (labeled by ID), colored by performance category. Performance categories were determined using a 0.5-point threshold for meaningful change. The horizontal dashed line at y=0 represents no change in performance.

Only two participants began with notably high baseline scores (≥4.0), both showing minimal change, possibly indicating a ceiling effect. Of particular interest is the observation that initial skill level did not completely determine improvement potential, as participants with similar baseline scores showed divergent improvement trajectories.

To triangulate SWIM performance with survey responses, we analysed three categories of responses:

1. Self-reported frequency of practicing the skill during the intervention ("behavioural experiments") was assessed in a survey at T2 with the item stem "Have you tried MI interaction techniques and other skills in your own work after training days 1 and 2?".

The response options were “I don’t know what this means”, “Not at all”, “With one service user”, “With a couple of service users”, “With many service users”, and “With all or almost all service users”.

2. Self-efficacy to use each skill was measured at T3 with an item stem “How certain are you, that you can use MI skills in your own work?”. Response options ranged from 1 (“Not at all certain”) to 5 (“Completely certain”).
3. Intention to use each skill was measured at T3 with the question “Do you intend to use the following MI skills and interaction techniques in your work?”, options being “Don’t intend to use at all”, “I intend to try”, “I intend to use occasionally”, “I intend to use relatively often”, and “I intend to use continuously”.

Figure S6 below presents a determinant category -level aggregation, where SWIM improvement groups are triangulated against survey questions. The individual data points show that while there is natural variation within each group, the overall distributions of scores support a theoretically grounded ordering, where improvers on the SWIM score also report higher self-efficacy, intention and practicing on the survey data.

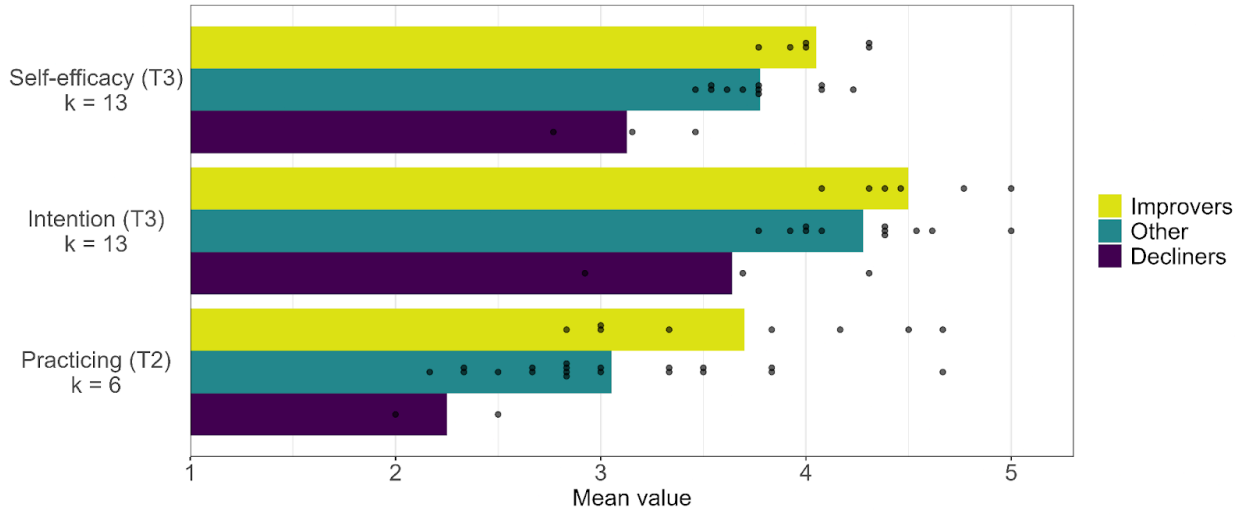

Figure S6. Means of self-reported behavioural correlates by performance groups based on SWIM coding (colours). Bars represent the aggregate mean score for each performance group (Improvers, Others, Decliners) within each composite variable. Points represent the mean score for each individual participant, per all k = 6 or 13 variables within the composite score.

Figure S7 below depicts results for each individual variable. If SWIM classifications were unrelated to survey responses, we would expect to see the hypothesized ordering (decliners lowest, improvers highest, others intermediate) in approximately 17 percent of variables (representing random arrangement). We can also observe that two practicing variables stand out: Few participants reported practicing change talk recognition or evocation.

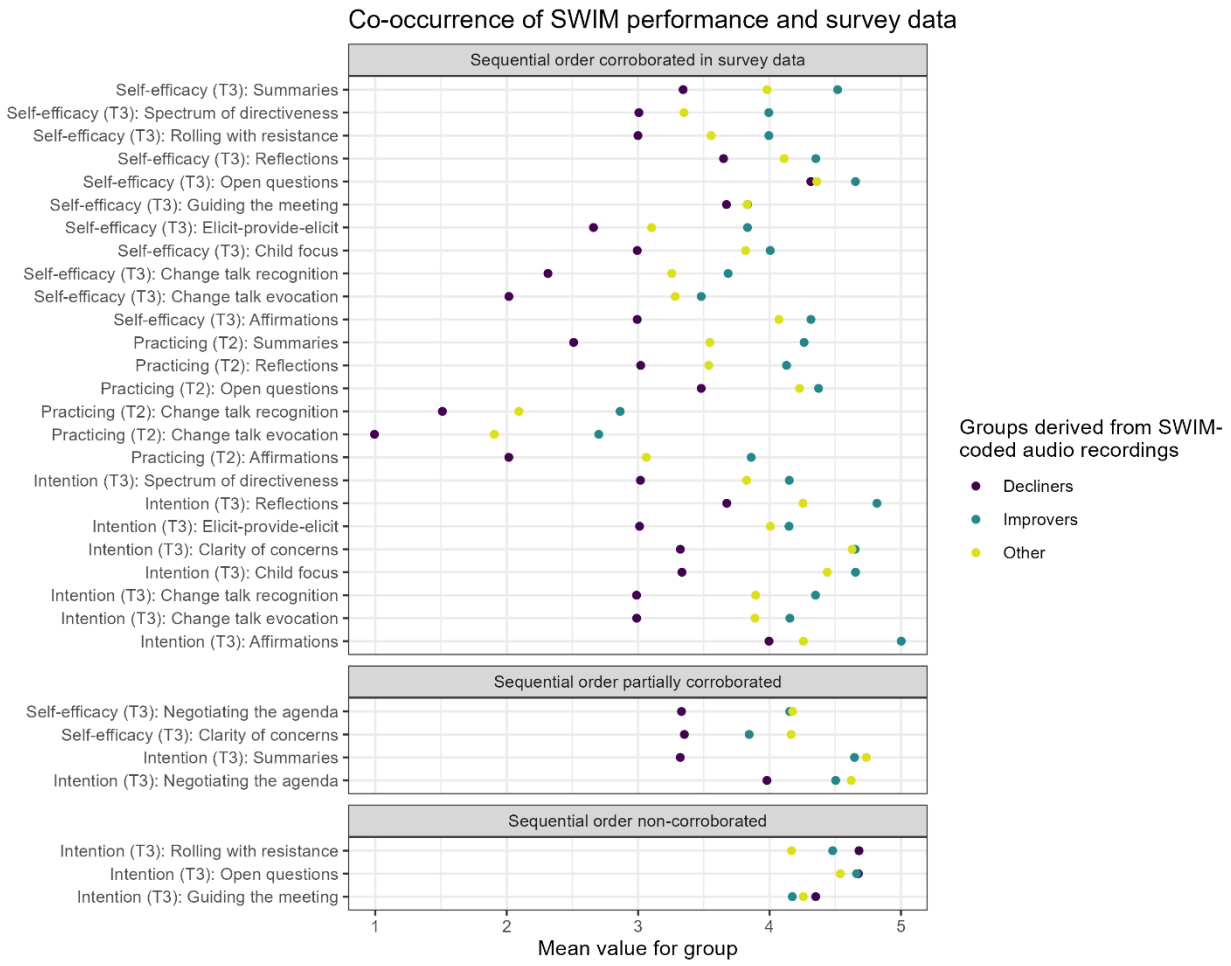

Figure S7. Comparison of mean survey responses for MI-related intentions, confidence, and practicing across SWIM-based performance groups (improvers, decliners, and others). Variables are grouped in panels, by degree of corroboration with SWIM-based classifications. Points represent group means on a 1-5 scale. We can observe that for most variables, the 'Improvers' category outperforms the 'Other' category, which in turn outperforms the 'Decliners' category. Change talk recognition and evocation stand out as exceptionally low scores within the Practicing-category.

We can also "zoom in" to comprehensively understand the variation within the aggregations of the figure, as represented in Figure S8 below.

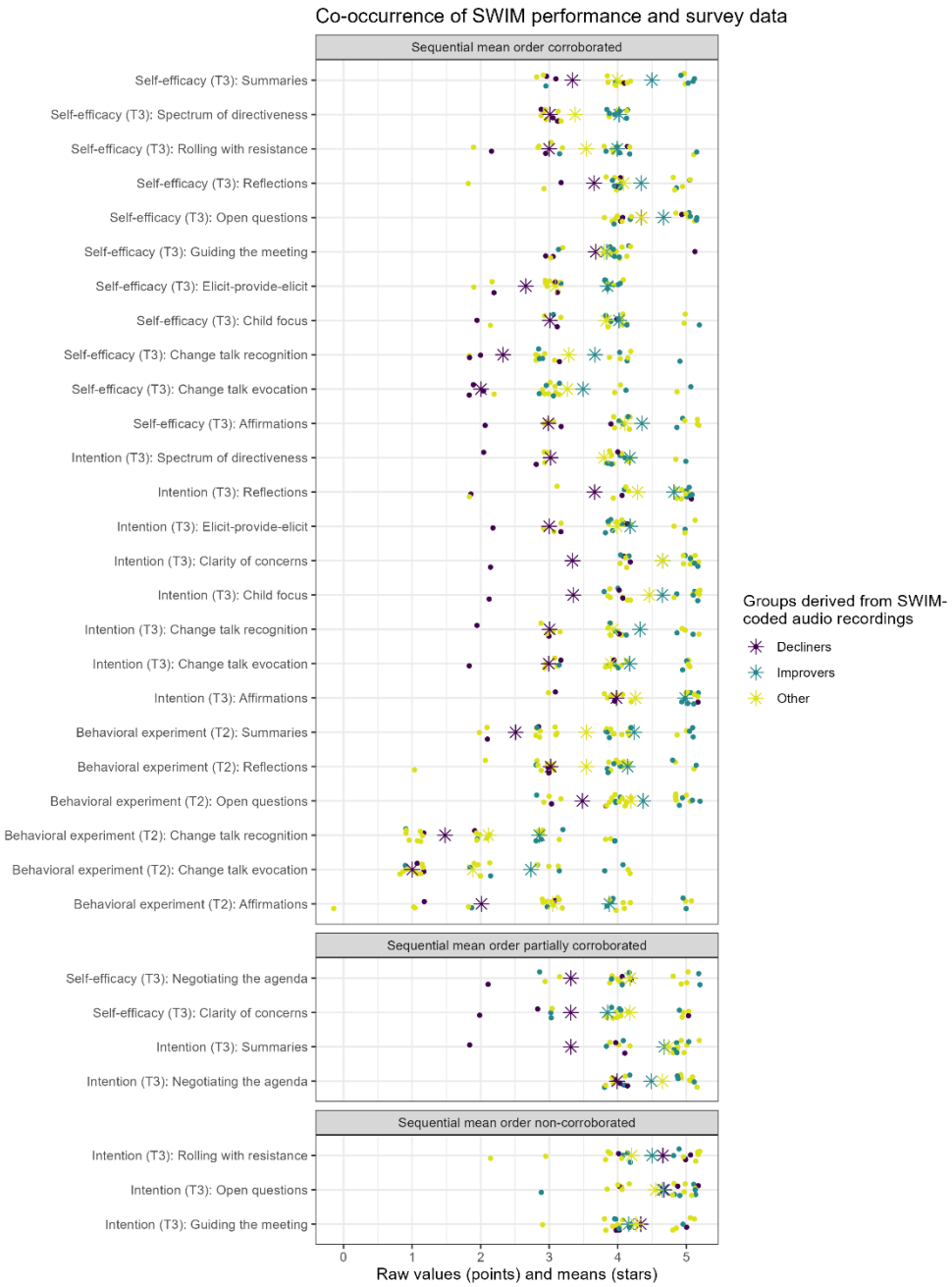

Figure S8. Raw data depicting variation around the means presented in the article's figure. Note: the single "outlier" in the decliner group, which seems to draw down the means, is not the same person for each variable.

3.2.2. Statistical procedures

To reduce the data and make it more interpretable, we composed new scales from each of the three survey response categories. Reliability and unidimensionality of these survey scales were assessed using McDonald's Omega. We assessed both total omega (ωt) for overall scale reliability and hierarchical omega (ωh) to determine the scale's unidimensionality (i.e., the extent to which it measures a single underlying construct).

The results showed that all three scales possessed good to excellent overall reliability (Self-Efficacy: ωt = 0.77; Practicing: ωt = 0.84; Intention: ωt = 0.89). Furthermore, hierarchical omega was high and nearly identical to total omega for all scales (Self-Efficacy: ωh = 0.77; Practicing: ωh = 0.84; Intention: ωh = 0.87), providing strong evidence that each scale measures a single, coherent construct and justifying their use as composite variables. Model fit statistics could not be reliably computed, which is common in analyses with small sample sizes and a forced single-factor solution. This does not invalidate the primary conclusion drawn from the strong ωh value regarding the scale's functional unidimensionality for the purposes of this study.

4. Intraclass correlation coefficient (ICC) (Table S14)

Average measure ICC (1,k) between two random raters.

Skill	ICC	p
Evocation	.661	.007
Collaboration	.812	<.001
Autonomy	.680	.002
Empathy	.754	<.001
Purposefulness	.778	<.001
Clarity about risk and need	.801	<.001
Child focus	.608	.010