Multimedia Appendix 9 – Kappa Reliability Scores of Credibility Checklist and Privacy Explanation Checklist Items

Inter-raters Agreement (Kappa) on Categorical Items that Assemble the Credibility Checklist

	Owner's Credibility	Maintenance	Strong Advisory Support	Evidence for Successful Implementation
Total (n=84)				
Kappa	.90	-[a]	.82	1.00
(95% CI)	(.81-1.00)	-	(.70-.94)	(1.00-1.00)
Mobile (n=42)				
Kappa	.87	.95	.79	1.00
(95% CI)	(.69-1.05)	(.85-1.00)	(.58-1.00)	(1.00-1.00)
Website (n=42)				
Kappa	.90	-[a]	.81	1.00
(95% CI)	(.76-1.03)	-	(.65-.96)	(1.00-1.00)
Mental Health (n=42)				
Kappa	.81	1.00[b]	.82	1.00
(95% CI)	(.63-.99)	(1.00-1.00)	(.67-.97)	(1.00-1.00)
Healthy Behaviors (n=42)				
Kappa	1.00	.90[b]	.76	1.00
(95% CI)	(1.00-1.00)	(.71-1.00)	(.51-1.00)	(1.00-1.00)

notes: Third Party Endorsement was not calculated since none of the programs was endorsed - reliability calculations are not applicable on constant measures.

[a] Maintenance was not evaluated for web-based programs and therefore reliability scores appear for mobile applications and separately by clinical aim. [b] Based on n=21 examined mobile applications.

Inter-raters Agreement (Kappa) on Categorical Items that Assemble the Privacy Explanation Checklist

	The system informs users of the data journey in detail to understand all sources of data exposure.	The system explicitly notifies how personal health information and/or personal identifiers will **be kept confidential** unless clear permission was given.	The system explicitly notifies how personal information and/or personal identifiers **may be used** before data is collected.	The system tunnels users through the terms of use explicitly.	The system allows users to keep identifiers private.	When not apparent the system lets users know when they go public.	The system warns the users from providing private information and asks permission to provide such information when applicable.
Total (n=84)							
Kappa (95% CI)	.95 (.88-1.00)	.98 (.93-1.00)	.95 (.89-1.00)	.70 (.46-.95)	1.00 (1.00-1.00)	1.00 (1.00-1.00)	.97 (.90-1.00)
Mobile (n=42)							
Kappa (95% CI)	.91 (.73-1.00)	1.00 (1.00-1.00)	.85 (.64-1.00)	.64 (.19-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
Website (n=42)							
Kappa (95% CI)	.95 (.84-1.00)	.94 (.82-1.00)	1.00 (1.00-1.00)	.73 (.44-1.00)	1.00 (1.00-1.00)	- [a] -	.95 (0.85-1.00)
Mental Health (n=42)							
Kappa (95% CI)	.95 (.84-1.00)	1.00 (1.00-1.00)	.90 (.77-1.00)	.69 (.41-.97)	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
Healthy Behaviors (n=42)							
Kappa (95% CI)	.95 (.85-1.00)	.95 (.86-1.00)	1.00 (1.00-1.00)	.66 (.03-1.00)	1.00 (1.00-1.00)	- [a] -	.94 (.83-1.00)

notes: [a] In these cases 100% agreement was noted, but Kappa was not calculated since all of these programs met criteria (constant result).

Interpretation

Based on criteria established by Landis and Koch (1977) the strength of agreement between raters was mostly at the outstanding agreement range (Kappa >.80; 44/51 ratings, 86.3%) with the minority of scores being at the substantial agreement range (.60 < Kappa < .80; 7/51 ratings, 13.7%).

Reference

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.