Headache
© 2020 The Authors. Headache: The Journal of Head and Face Pain published by Wiley Periodicals, Inc. on behalf of American Headache Society.
ISSN 0017-8748
doi: 10.1111/head.13772
Published by Wiley Periodicals, Inc.

Research Submission

Biofeedback Treatment App for Pediatric Migraine: Development and Usability Study

Anker Stubberud, MD ; Erling Tronvik, PhD ; Alexander Olsen, PhD ; Gøril Gravdahl, MSc; Mattias Linde, PhD

Objective.—The objective of this study was to develop and investigate the usability of a biofeedback treatment smartphone app for adolescent migraine sufferers.

Background.—Biofeedback is effective in treating pediatric migraine. However, biofeedback is not widely used due to the necessity of a trained therapist and specialized equipment. Emerging digital technology, including smartphones and wearables, enables new ways of administering biofeedback.

Methods.—In a prospective open-label development and usability study, 10 adolescent migraine sufferers used a newly developed biofeedback app with wearable sensors that measured their muscle tension, finger temperature, and heart rate. Three iterative rounds of usability testing, including a 2-week home testing period, were completed. A biofeedback algorithm, combining and optimizing the 3 physiological modalities, and several algorithms for sham-treatment were created. Usability was evaluated statistically and summarized thematically.

Results.—Five of ten participants completed all 3 rounds of usability testing. A total of 72 biofeedback sessions were completed. Usability scoring was consistently high, with median scores ranging from 3.5 to 4.5 on a 5-point scale. The biofeedback optimization algorithm correlated excellently to the raw physiological measurements ($r = 0.85$, $P < .001$). The intervention was safe and tolerable.

Conclusion.—We developed an app for young migraine sufferers to receive therapist-independent biofeedback. The app underwent a rigorous development process as well as usability and feasibility testing. It is now ready for clinical trials.

Key words: mHealth, smartphone, wearables, headache, adolescent

Abbreviation: mHealth mobile health

(*Headache* 2020;60:889-901)

From the Department of Neuromedicine and Movement Science, NTNU Norwegian University of Science and Technology, Trondheim, Norway (A. Stubberud, E. Tronvik, G. Gravdahl, and M. Linde); National Advisory Unit on Headaches, Department of Neurology and Clinical Neurophysiology, St. Olavs Hospital, Trondheim, Norway (E. Tronvik, G. Gravdahl, and M. Linde); Department of Psychology, NTNU Norwegian University of Science and Technology, Trondheim, Norway (A. Olsen); Department of Physical Medicine and Rehabilitation, St. Olavs Hospital, Trondheim, Norway (A. Olsen).

Address all correspondence to A. Stubberud, Department of Neuromedicine and Movement Science, NTNU Norwegian University of Science and Technology, Trondheim, Norway, email: anker.stubberud@ntnu.no

Accepted for publication January 24, 2020.

INTRODUCTION

Pediatric migraine is highly prevalent and associated with the substantial deterioration of social functioning and mental health.[1,2] There are few viable options for prophylactic medications, with many options having limited efficacy or adverse effects.[3,4] However, behavioral prophylaxis appears to be a valid treatment option for pediatric pain and headache.[5,6] Specifically, biofeedback is one of the most prominent behavioral approaches, and meta-analytical evidence suggests that it is effective in treating pediatric migraine.[7]

Despite being effective, biofeedback has limited population coverage. This is possibly because it is time-consuming and costly with its provision traditionally through specialist clinics. Typically, to be effective, biofeedback treatment requires a trained therapist, as well as specialized equipment measuring surface electromyography, peripheral skin temperature, or heart rate.[8] However, new digital technologies, including wearable sensors and the use of smartphones for medical purposes mobile health (mHealth), provide new possibilities.[9] Recent research suggests that behavioral mHealth interventions for headache are feasible, but development processes and usability testing remain insufficient.[10] Additionally, efficacy measures are uncertain.[11] Currently, there are no biofeedback smartphone applications available specifically targeted at pediatric migraine.[12] To address this, we have recently performed a study showing that wearable sensors are suitable for biofeedback,[13] similar to studies that have validated the use of wearables for other medical purposes.[14,15] Nonetheless, mHealth treatment is entirely dependent on robust development and usability testing to ensure adherence and efficacy.[10,16]

We present a development and usability study aimed at (1) developing a new biofeedback app for adolescents with migraine and evaluating and improving its feasibility and usability; (2) developing and optimizing an algorithm for the multimodal combination of data from selected physiological measures to provide personalized and therapist-independent biofeedback; and (3) developing a sham biofeedback paradigm to be used as a control in efficacy trials.

METHOD

Study Design and Participants.—The study was designed as a prospective open-label iterative and incremental development and usability study at St. Olavs University Hospital in Trondheim, Norway, from September 2017 to June 2018. Ten adolescent migraine sufferers (aged 13-17 years) were recruited from the municipality using social media and the hospital intranet. No statistical power calculation was conducted prior to the study, and the sample size was based on recommendations for usability studies. All diagnoses were confirmed by a consultant neurologist with headache expertise. The participants completed 3 cycles of usability testing with a smartphone biofeedback app. The first two were conducted in a makeshift usability lab, while the final cycle was performed over 14 days

Conflict of Interest: NTNU and St. Olavs Hospital, Trondheim University Hospital may benefit financially from the commercialization of the proposed treatment through future possible intellectual properties. This may include financial benefits to the authors of this article. Dr. Stubberud is a co-founder of the Nordic Brain Tech AS, a spin-off company that was established based on the proposed treatment in this and previous studies at the NTNU Norwegian University of Science and Technology. Dr. Stubberud is a co-inventor of the proposed treatment in this study and may benefit financially from a license agreement between Nordic Brain Tech AS and NTNU Norwegian University of Science and Technology. Dr. Tronvik is a co-founder of the Nordic Brain Tech AS, a spin-off company that was established based on the proposed treatment in this and previous studies at the NTNU Norwegian University of Science and Technology. Dr. Tronvik is a co-inventor of the proposed treatment in this study and may benefit financially from a license agreement between Nordic Brain Tech AS and NTNU Norwegian University of Science and Technology. Dr. Olsen is a co-founder of the Nordic Brain Tech AS, a spin-off company that was established based on the proposed treatment in this and previous studies at the NTNU Norwegian University of Science and Technology. Dr. Olsen is a co-inventor of the proposed treatment in this study and may benefit financially from a license agreement between Nordic Brain Tech AS and NTNU Norwegian University of Science and Technology. Mrs. Gravdahl declares no potential conflicts of interest concerning the research, authorship, or publication of this article. Dr. Linde is a co-inventor of the proposed treatment in this study and may benefit financially from a license agreement between Nordic Brain Tech AS and NTNU Norwegian University of Science and Technology.
Funding: The study received funding for cooperative projects between the Department of Neuromedicine and Movement Science and Department of Psychology NTNU Norwegian University of Science and Technology.
Registration: The study was approved by the regional committee for medical and health research ethics (2017/582-3) and registered at the Norwegian Centre for Research Data (project number: 54571) prior to enrollment of participants.

at home. After the usability testing, the data collected were used to develop an algorithm to process and combine multimodal physiological data for biofeedback, develop an algorithm for sham-treatment, and finalize the app design to be used and further tested in clinical trials. The study was approved by the regional committee for medical and health research ethics (2017/582-3) and the Norwegian Centre for Research Data (project number: 54571).

Inclusion criteria were age between 12 and 18 years; migraine with or without aura (MWA or MWoA) diagnosed according to the International Classification of Headache Disorders 3;[17] 2 to 6 attacks per month; not using prophylactic migraine medication; experience with using an iPhone® (Apple Inc.); and informed signed consent provided by their guardian. Exclusion criteria were lack of proficiency in the Norwegian language; reduced vision, hearing, or sensibility to a degree that hampered study participation; or if they had any serious neurological or psychiatric disorders.

Biofeedback Setup.—The biofeedback setup consisted of 3 sensors measuring muscle tension, finger temperature, and heart rate, all transmitting signals via Bluetooth® Smart/4.0 to an iPhone® 6 or newer. A small compact bipolar surface electromyography sensor (NeckSensor™; EXPAIN AS, Oslo, Norway) was used for measuring muscle tension from the upper trapezius muscle fibers. A PASPORT Skin/Surface Temperature Thermistor Probe, PS-2131 (Pasco, Roseville, CA, USA) was held between the index finger and thumb of the right hand to measure finger temperature. Finally, a MIO Fuse™ (Mio Global, Physical Enterprises) heart rate wristband was used to measure heart rate over the dorsal aspect of the left wrist.

Usability Evaluation and App Development.—Usability evaluation and biofeedback app development consisted of 3 iterative cycles. Each cycle included the following steps: (1) app programing and design; (2) intervention review by a neurologist, neuropsychologist, computer engineer, and medical student; and (3) usability testing by adolescent migraine sufferers. The initial version of the app-user interface was based on a literature review and evaluation from a previous study that validated wearable sensors as suitable for biofeedback.[13]

The first two usability-testing cycles were completed as one-hour sessions in a consultation room at the hospital. During the first cycle, the participants were initially given an introduction, a description of the rationale of the treatment, and instructions on how to use the app. Subsequently, they were asked to set up the equipment, start the app, and complete a ten-minute biofeedback session. Participants were not trained or instructed in relaxation or stress management techniques. For the second cycle, the participants completed 3 sessions of 5 minutes, with 20 minutes rest time between each session. The final cycle was conducted at home for 2 weeks. The participants were provided with sensors to be used with their personal iPhone®. They downloaded the app from a webpage and were asked to complete a daily biofeedback session of 10 minutes. Following this, they completed a headache diary in the app. After each usability cycle, the participants were asked to complete a comprehensive, structured age-appropriate user evaluation (Supporting Information 1). The user evaluation form was based on commonly structured surveys such as the *Post Study System Usability Questionnaire*, the *System Usability Survey*, and a recently developed mobile app rating scale.[18] The 5 main domains included in the evaluation were (1) engagement; (2) functionality; (3) design; (4) information; and (5) understanding of the biofeedback. The user evaluation also included questions regarding any discomfort they experienced while using the app or sensors and an open-ended adverse events assessment. During the 2 first sessions, 1 of the investigators was present to assist participants with completing the evaluation. Experiences and findings from the intervention review and usability testing from each cycle were used to implement changes to the app for the next iteration of testing. Descriptive analyses of changes to the app interface and development were summarized by a simple thematic analysis categorized under the same 5 domains as the questionnaire.[18]

Biofeedback Algorithm Development.—The biofeedback algorithm was designed to give a compound feedback signal based on all 3 input parameters, that is, muscle tension, finger temperature, and heart rate. To optimize feedback, 2 settings of the algorithm were individually adjusted to each user. First, the default upper and lower measurement limits for the 3 physiologic

parameters were defined based on normalizing graphs of participant data. A factor was then defined as to how the upper and lower limits would be adjusted between each session. Based on the upper and lower individual physiological limits, a 0-100 score for each parameter was created. Second, we defined an internal weighting factor for combining the 3 parameter scores. This was to ensure that a lack of improvement in 1 parameter for a session and absence of a decreasing score would still result in a moderate positive combined score. These variable factors were decided based on the usability evaluation and confirmed as suitable using a regression analysis after the final iteration.

We also developed a set of sham-algorithms by manipulating the raw data. The sham algorithms were visually and statistically analyzed to evaluate if they produced sufficient disruption between the physiological data and feedback, while, importantly, still retaining masking and motivation for the user.

Data Management and Statistics.—The average number of hours of daily smartphone use, general experience with apps, and experience with wearable sensors were averaged over the 3 cycles for each participant. Usability evaluations were scored on a 5-point Likert interval scale, ranging from 1-"completely disagree" to 5-"completely agree." These scores were averaged over each domain for all participants. We used the principle of last observed value carried forward (LOCF) for missing data from dropouts in the usability analyses. We also made an analysis of complete data to serve as a comparison to the imputed data. Baseline feedback score and change in feedback score (ie, the change from the start to the end of a session) for surface electromyographic voltage, skin temperature, and heart rate were registered for all completed sessions. Combined unweighted "raw" scores were created using an equal 33.3% weighting for each of the 3 physiologic parameter scores, while biofeedback algorithm weighted change values were calculated using the above-described biofeedback algorithm. We used only complete data for analyses of physiological measurements without imputing data.

Data were reported as means, standard deviations (SD), medians, and interquartile ranges (IQR). Usability scores were compared between cycles with a two-tailed Wilcoxon signed-rank test and summarized with medians and IQR. We calculated the Pearson correlation coefficient to assess the association between the combined unweighted scores and biofeedback algorithm scores and described the association using a two-tailed linear regression analysis. The regression analysis was applied to evaluate if the biofeedback algorithm would provide a non-random and systematic improvement in feedback scores. All normality assumptions were checked by visual inspection of histograms. P values $<.05$ were considered statistically significant.

This is the primary analysis of data collected in this study. A priori we planned for analyses to compare scores across usability cycles and analyze for correlation between the raw feedback scores and the algorithm scores. Analyses of correlation between familiarity with apps and wearables and usability scores were also planned a priori but were omitted as data were underpowered and not suited for regression analyses.

All statistical analyses were performed and figures were made using Stata v14 (Stata Corp, College Station, TX, USA) and Python v3.6 (Python Software Foundation) with the pandas v0.20.3, NumPy v1.17.2, matplotlib v3.1.1, and scikit-learn v0.21.3 libraries.

RESULTS

Participants and Demographics.—Ten participants with a mean age of 15 ± 1.6 years (range, 13-17 years) were included in the study. Seven were male. One participant did not attend the first cycle. In the second cycle, 2 dropped out, and 1 did not attend. In the final cycle, 2 additional participants dropped out, and 1 had problems with making the setup work properly. Five participants completed all usability cycles and 5 of 10 participants dropped out (50% attrition rate). The average daily previous smartphone usage was 3.7 ± 1.6 hours. The median value familiarity with previous smartphone apps was 4 (good familiarity), with a mean of 4.0 ± 0.8, while the median value familiarity with wearable sensors was 1 (very little familiarity), with a mean of 1.5 ± 1.0. A total of 72 biofeedback sessions were completed throughout the study, with an average per participant of 8.4 in the 2 weeks of the third cycle.

Usability Metrics and App Development.—Figure 1 shows the median and IQR usability scoring for the 5 primary domains of usability assessment

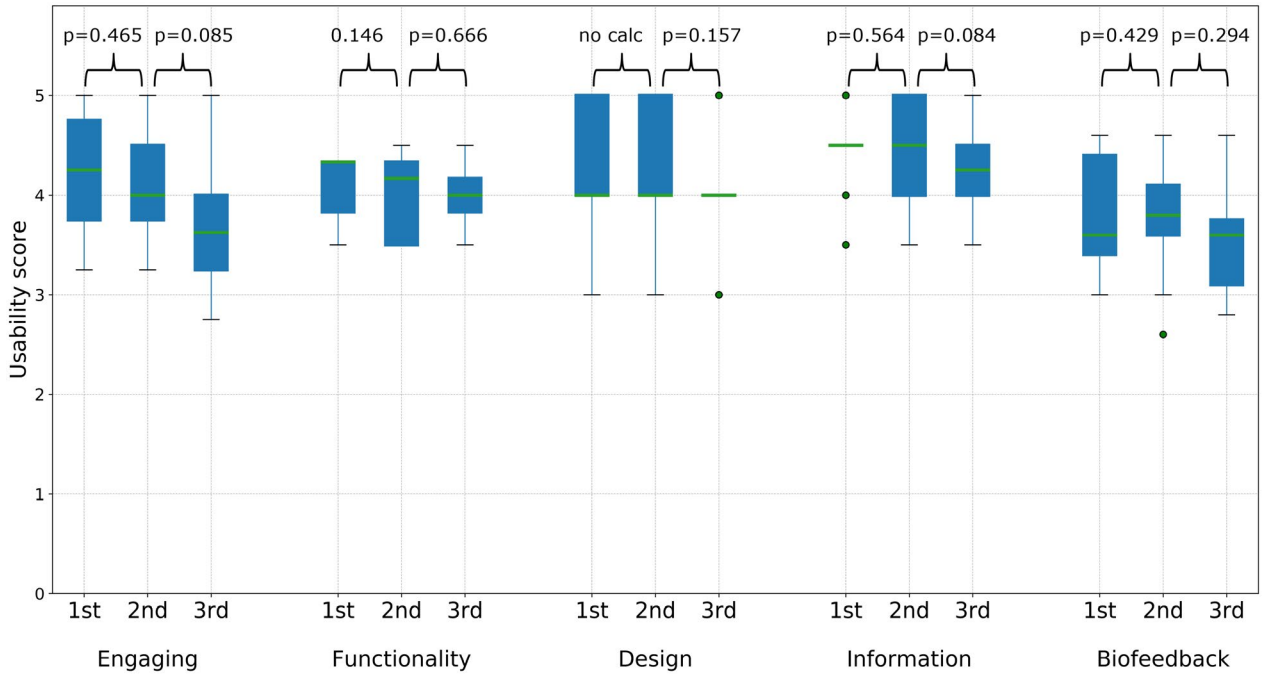

Fig. 1.—Boxplot showing the usability scoring for the 5 main domains through the 3 usability cycles. Green horizontal lines represent medians, blue boxes represent inter quartile ranges (IQR), whiskers represent IQR × 1.5, and green dots represent outliers. Usability scores were compared between cycles with a two-tailed Wilcoxon signed-rank test. No statistically significant changes were found between iterations for each of the 5 domains. Test statistics and *P* value for the domain of design from first to second cycle was not calculable because all ranks were tied. [Color figure can be viewed at wileyonlinelibrary.com]

Table 1.—Thematic Summary of the Most Important App Development and Interface Changes Implemented After Each Cycle. Themes are Classified According to the 5 Main Usability Domains[18] – Biofeedback, Design, Engaging, Functionality, and Information

Cycle 1	Biofeedback – Feedback as separate visualizations for each parameter instead of a combined circle implemented
	Design – Bright light, and especially bright blue colors avoided
	Functionality – Enabled easier navigation and flow between app screens
Cycle 2	Biofeedback – Feedback sometimes perceived as too sensitive was thereby smoothed over a short window
	Engaging – Included reminder function at set timepoint daily
	Information – Provided better information on how to use and connect sensors in the app
Cycle 3	Biofeedback – Algorithm for weighted and individualized feedback optimized
	Design – Finished design with a desirable color palette of dark and green
	Functionality – Included interactive diary that easily allows for viewing previous sessions and headache entries

for each of the 3 iterations. No statistically significant difference was found from the first to the second cycle and from the second to the third cycle for any of the domains (Fig. 1). The complete data analysis without LOCF imputation resulted in lower estimates in the third cycle domains of engaging and biofeedback, with medians (IQR) at 3.3 (2.9-3.6) and 3.2 (3.0-3.5), respectively. Thematic descriptions of the changes in the app interface and app development implemented after each cycle are provided in Table 1. Figures 2 and 3 depict the app interface before the first iteration and after the final iteration.

Biofeedback Algorithm and Sham Development.—The mean value ±2 standard deviations was used to establish the default upper and lower measurement limits for all physiologic parameters. Out of the 251,874 data points for muscle tension measurements, 95% of the values fell within a range of 0.01-0.16 mV.

Fig. 2.—Sample of screenshots from the first version of the app. Screenshot 1 shows the 3 physiological parameters combined as 1 feedback visualization. Screenshot 2 shows the first edition of the headache diary with a subscreen or roll-down pane for each question. [Color figure can be viewed at wileyonlinelibrary.com]

Fig. 3.—Sample of screenshots from the final version of the app. Screenshot 1 shows the easily navigable headache diary. Screenshot 2 shows the feedback as 3 separate feedback-indicators, 1 for each physiologic parameter. Screenshot 3 shows one of the pictures for instructions on connecting sensors. Screenshot 4 shows the headache diary overview. [Color figure can be viewed at wileyonlinelibrary.com]

Similarly, from the 18,572 data points for heart rate measurements, 95% of the values fell within a range of 46 to 90 beats per minute. Finally, out of the 20,734 data points for the finger temperature measurements, 95% of the values fell within a range of 25.3-39.0°C. This supernormal upper-temperature limit is caused by uncertainty in the absolute measurements of the temperature sensor. Therefore, the default upper limit was set to 37.0°C and was allowed to vary around this limit. The default lower temperature limit was set according to the 2.5‰.

Data from 42 completed biofeedback sessions in the third cycle were used to calculate "raw" unweighted scores and biofeedback algorithm scores on a 0-100 scale. The unweighted baseline score was 64.1 ± 10.6 and the end-session unweighted scores were 72.0 ± 10.3. Applying the biofeedback algorithm to the same dataset yielded a baseline session score of 64.3 ± 10.6 and an end-session score of 78.5 ± 10.7. A Pearson's product-moment correlation analysis established a strong positive correlation between the change in unweighted and biofeedback algorithm scores, $r(40) = 0.85$, $P < .001$. The corresponding linear regression established that the unweighted scores accounted for 72% of the variation in the crude biofeedback algorithm scores with the following regression equation: biofeedback algorithm scores = 7.41 + 0.85 × (unweighted score), $F(1, 40)$, $P < .001$. Figure 4 is a scatter plot showing the regression line of fit to visualize the linear correlation and illustrate how the biofeedback algorithm results in an improved feedback score, whereas a sham-algorithm leads to random feedback scores.

Four principal approaches were attempted to develop sham biofeedback. These are described in detail and evaluated in Table 2 and Figures 4 and 5. Sham biofeedback, where the feedback is distorted by a sine-wave fluctuation, was considered the most suitable. This sham was judged to give incorrect feedback, but not to the degree that would promote unmasking.

Safety and Tolerability.—From the evaluation questionnaires, 12 out of the 20 ratings relating to intervention discomfort were "very little discomfort," while the remaining 8 were "little discomfort." Out of the 20 ratings relating to sensor discomfort, 14 were "very little discomfort," 5 were rated as "little discomfort," and 1 was rated as "very great discomfort." No serious adverse events were reported.

DISCUSSION

Principal Findings.—We developed a new mHealth biofeedback intervention for young migraine sufferers that is suitable for self-administration. The intervention includes an algorithm that gives optimized and personalized compound feedback based on 3 physiological parameters proven to be effective in migraine prophylaxis. The intervention was perceived as safe and received consistently high usability scores throughout the 3 cycles of usability testing.

Interpretation.—We developed an app with an algorithm that combines 3 physiological parameters, as opposed to traditional biofeedback where 1 parameter is used.[19] This optimization algorithm was implemented to overcome the challenge that not all biofeedback users experience an influence over the physiological parameter measured,[20] and that different parameters may be useful for different users. For instance, if a user excels at raising their finger temperature, but has trouble lowering their heart rate, the algorithm will fade out the latter throughout the session and thereby chose a more appropriate and "personalized" parameter for the individual. Comparably, the parameter that is most efficient for each user will be given the heaviest weighting in the combined feedback score. This feature was implemented believing that it is likely to result in relevant and useful feedback for a larger group of potential users. Moreover, the intervention did not include commonly used adjuvant therapies such as relaxation training and stress management techniques. This was a deliberate decision made to investigate both if a therapist may be completely excluded from the usual biofeedback treatment "package," and to see if the app itself may to a certain degree replace the therapist. The algorithm was also deemed as suitable after a regression analysis, where the algorithm yielded systematically improved scores, with a significant proportion still attributable to the raw data. This confirmed the desired effect of the biofeedback algorithm to give moderate positive combined feedback despite lack of continuous "improvement" in a physiological parameter. Additionally, the app enables personalized scoring of physiological parameters in an age group that is

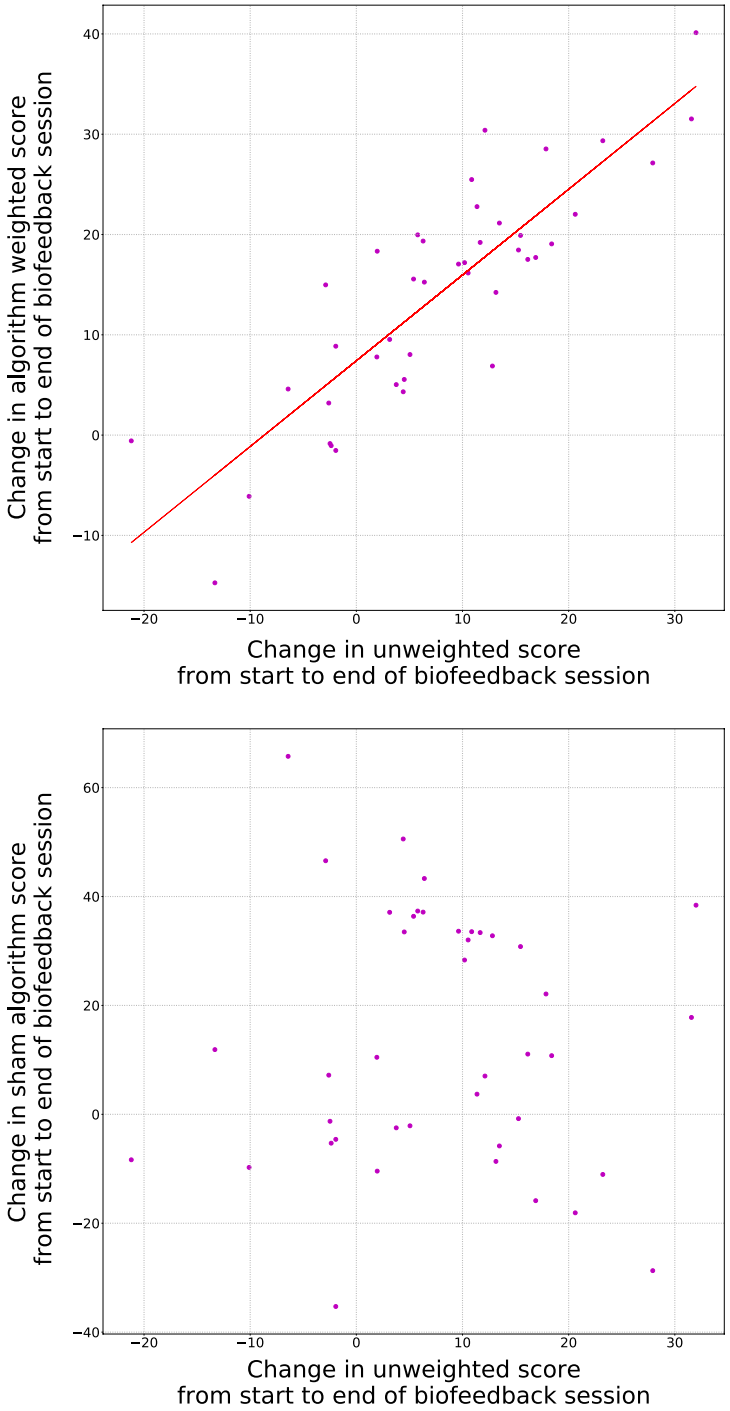

Fig. 4.—Scatterplots comparing "raw" unweighted feedback scores to the biofeedback algorithm scores (upper plot) and sham scores (lower plot) based on 42 biofeedback sessions completed by the 5 participants in the third usability cycle. The values on the axes represent change in feedback score from the start to the end of a biofeedback session (ie, a user's performance during that session). The upper scatterplot shows biofeedback algorithm change values plotted against unweighted change values. The biofeedback algorithm change values are generally higher than the unweighted change values indicating that the biofeedback algorithm improves feedback scores. The predicted improvement in scores is given by the regression equation (red line) for the linear regression model: Biofeedback algorithm score = 7.41 + 0.85 × (unweighted score), $F(1, 40)$, $P < .001$. Together, this illustrates that the biofeedback algorithm scores are improved while still preserving a relationship to the actual "raw" unweighted data. The lower scatterplot shows the inverted sham algorithm change values plotted against unweighted change values. Contrary to the biofeedback algorithm, there is no clear relationship between the sham change values and the unweighted change values. This sham gives a very random feedback and was thus deemed as unsuited because it would likely promote unmasking. [Color figure can be viewed at wileyonlinelibrary.com]

Table 2.—Sham Biofeedback Alternatives

Sham Name	Description	Evaluation
Inverted weighting	Inverting each parameter score and weighting of the 3 physiologic parameters	Applying the inverted weighting algorithm to the raw data yielded a baseline session score of 36.4 ± 12.4, and an end-session score of 50.3 ± 19.4. Moreover, it produced a nearly random feedback score with no clear relationship to the unweighted scores (Fig. 4) and was thus deemed unsuited
Sine-wave fluctuations	Applying a sine-wave fluctuation multiplier of amplitude a to the raw combined data: Sham = $\sin(r \times w \times a)$	The sine-wave fluctuation was evaluated at wavelength (w) 0.1, 0.05, and 0.01π; and at amplitudes (a) 0.05, 0.10, 0.15, and 0.20. The sine-wave fluctuations produced a sham signal deemed to be sufficiently disrupted form the raw data, but still not giving obvious signal deviations in cases such as voluntary contractions and loss of sensor contact. The most suited sine-wave sham version is visualized in Figure 5
Random fluctuations	Applying a pseudo-random fluctuation multiplier of amplitude a to the raw combined data: Sham = $r \times$ random $\{\text{random} \in \mathbb{R} \mid 1-a \leq \text{random} \leq 1+a\}$	The pseudo-random fluctuation multiplier was evaluated at frequencies 1, 0.5, 0.33, 0.25, and 0.1 Hz; and at amplitudes (a) 0.10, 0.15, and 0.25. The random fluctuations were evaluated as producing sufficient disruption of the signal, but to a degree that might promote unmasking, and thus deemed unsuitable. The most suited random fluctuation sham is visualized in Figure 5
Full disruption	Providing a feedback signal completely separated from the actual physiological measurements	A full disconnection between the input physiologic data and the feedback visualization, for example, by presenting a completely random feedback, was evaluated as unsuited because it would easily lead to both unmasking and demotivation with the user

known to display great variance in their physiological properties.[21,22] Together, this provides robust therapist-independent treatment.

In addition to the development of the biofeedback algorithm, we rigorously tested and evaluated several sham-treatments. An empirical evaluation of the sham algorithms would have been beneficial to accurately ascertain what type of sham would best perform in a controlled trial. However, this was not prioritized in this current study as our main aim focused on usability. Nevertheless, this paper presents several potential shams of which both random fluctuation and sine-wave fluctuation shams were considered suitable. The inverted weighting and full disruption shams should be avoided because they may promote unmasking of the sham control.

The intervention was considered tolerable and safe. One participant reported "very great discomfort" after using the intervention. This is most likely due to these questions having inverted scoring as compared to the majority of questions in the evaluation. We have previously captured experiences of unpleasantness when removing the electromyography electrodes,[13] but this was not the case in this study. Finally, serious adverse events were not expected and have not been reported in the literature.[7,19] No serious adverse events occurred during our study.

Throughout this study, we aimed to assess and improve the feasibility and usability of the app, which is essential to obtaining satisfactory adherence and an effective treatment.[23,24] Such a rigorous usability approach yields important results that are highly informative for further development, and critical for planning clinical trials. It may be considered as similar to the phase I-II development of new drug treatments.[25] Similar studies carried out within other medical fields have also detected and addressed several issues regarding the feasibility and usability of mHealth interventions. Among these, several[26,27] also used an iterative approach, which is an established usability strategy.[28] Altogether, this highlights the necessity of development and usability studies when creating mHealth interventions. In our study, the usability scorings were consistently high. We evaluated the effect of changes

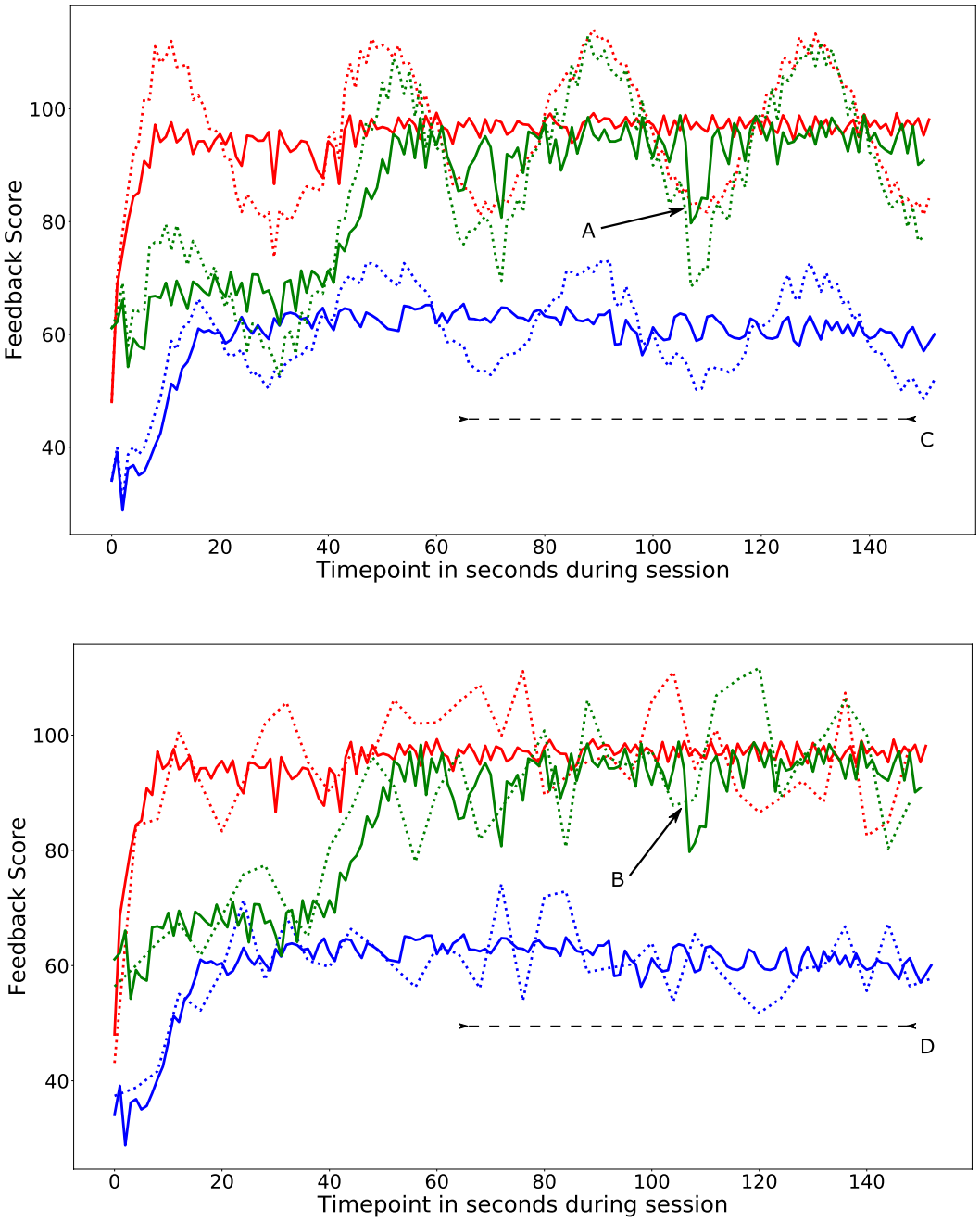

Fig. 5.—Lineplots with the "raw" unweighted feedback scores (solid lines) and sham scores (dotted lines) for 3 representative participants by color. The upper plot shows the sine-wave fluctuation sham scores (dotted lines) with a wavelength of 0.05π and a multiplication amplitude of 0.15. The lower left plot shows the random fluctuation sham scores (dotted lines) with a frequency of 4 Hz and a multiplication amplitude of 0.15. The figures are intended to illustrate how the sham feedbacks are experienced by the user, as compared to the "raw" unweighted feedback scores. Arrow A and B points to a timepoint in the green participant's biofeedback session with a sudden drop in "raw" feedback scores as may occur upon shrugging the shoulders (sudden increase in electromyographic voltage) or losing contact with the finger heart rate sensor (sudden fall in heart rate). In the upper plot, arrow A points to a corresponding decrease in the sine-wave fluctuation sham score. In the lower plot, arrow B points to a moment where the random sham results in a sudden increase in feedback scores, despite an obvious drop in the "raw" feedback score. Such randomness might promote unmasking, thus making a random fluctuation sham less suited. Moreover, the dotted horizontal lines C and D represent a time period where the "raw" feedback score for the red and blue participant is relatively stable. During this time period, the sine-wave fluctuations gives incorrect feedback which is slow and smooth, whereas the random fluctuation gives sharp and sudden changes in feedback score further promoting unmasking. [Color figure can be viewed at wileyonlinelibrary.com]

in each iteration on the usability scores, but no systematic statistically significant differences between cycles were found. This may be explained by both the original high scores and the sample size not providing sufficient power to detect a difference. On the contrary, the high attrition rate should also be considered as a measurement of usability, and dropping out of the study may simply be the result of a participant not enjoying the app. Likewise, missing data as a result of attrition certainly impacts interpretation of usability scoring. If all dropouts were in fact not liking the app, the overall usability scoring would have been poorer. In addition to the quantification of the usability, the participants were asked several open-ended questions and interviewed during the evaluation. Their comments provided valuable qualitative input for a thematic analysis on how the app could be improved. The app-user interface and usability were qualitatively improved with each iteration even though this was not evident in the usability scoring. We believe this resulted in a final solution that is more likely to meet the desired needs of a larger group of users as compared to an undocumented product directly being implemented in clinical efficacy trials.

Limitations and Strengths.—Several factors limit this study and make us reluctant to draw firm conclusions. First, the questionnaire was not validated for our specific study but rather based on common usability surveys and validated mHealth questionnaires.[29] Such questionnaires can be susceptible to response bias,[30] including acquiescence bias, in which participants automatically endorse statements to please the interviewer.[31] This may explain the high usability scorings in the initial cycle and the lower scores in home testing. Second, the first two usability cycles were conducted in a controlled environment, not fully representative of the intended use. In addition, the home testing session was conducted over a shorter than recommended period.[32,33] Together, this adds some uncertainty concerning the adherence to the intervention. Third, the study had a moderate sample size and suffered from attrition. This may represent poor usability and decrease the confidence in our findings. Nonetheless, the sample size was chosen according to recommendations for usability studies. Some researchers even argue that a sample size of approximately 5 participants is sufficient to uncover the majority of usability problems,[34,35] while others argue that such a small sample size is insufficient and that sample sizes should be customized to individual studies.[36] We ultimately chose a sample size of 10 people stratified across the adolescent age range to ensure essential usability problems were uncovered, while also receiving an evaluation from the whole heterogeneous age spectrum.

This is the first study of adolescent migraine that uses mHealth to deliver migraine therapy and enables biofeedback treatment to be provided to a broader population. The optimizing algorithm included in the intervention makes it superior to traditional monitoring that requires a trained therapist for interpretation. This will, in turn, lower costs and increase availability. Moreover, the intervention was developed by a multidisciplinary team, including neurologists with headache expertise, a neurophysiologist, and software engineers based on the guidelines for developing mHealth apps[37] and guidelines for behavioral treatment trials.[33] It used sensors that have previously been validated as appropriate for biofeedback and we involved the target group throughout the whole development process. These factors all helped to improve the final product.[23,38] By using the same set of participants for all cycles of usability testing, we also overcame the challenge that biofeedback as a psychophysiological training method requires several rounds of exposure to master.[8] This also allowed us to complete a large number of biofeedback sessions and repeated usability tests on the same individuals. Altogether, we believe that this new intervention has the potential to be effective and reach a broader population in need.

CONCLUSION

In this study, we developed a new biofeedback treatment app targeted at young migraine sufferers. The treatment includes wearable sensors, validated as appropriate for biofeedback, and a feasible and usable app developed specifically for the target population. Some study findings were limited by the low sample size, attrition, and response bias. Future studies should determine whether the migraine intervention developed in this study has a clinical effect on the migraine burden in adolescents.

Acknowledgments: The authors thank all participants for taking part in the study.

STATEMENT OF AUTHORSHIP

Category 1

(a) Conception and Design
Anker Stubberud, Erling Tronvik, Alexander Olsen, Mattias Linde

(b) Acquisition of Data
Anker Stubberud, Erling Tronvik, Gøril Gravdahl, Mattias Linde

(c) Analysis and Interpretation of Data
Anker Stubberud, Erling Tronvik, Alexander Olsen, Mattias Linde

Category 2

(a) Drafting the Manuscript
Anker Stubberud

(b) Revising It for Intellectual Content
Anker Stubberud, Erling Tronvik, Alexander Olsen, Gøril Gravdahl, Mattias Linde

Category 3

(a) Final Approval of the Completed Manuscript
Anker Stubberud, Erling Tronvik, Alexander Olsen, Gøril Gravdahl, Mattias Linde

REFERENCES

1. Wober-Bingol C. Epidemiology of migraine and headache in children and adolescents. *Curr Pain Headache Rep*. 2013;17:341.
2. Krogh AB, Larsson B, Linde M. Prevalence and disability of headache among Norwegian adolescents: A cross-sectional school-based study. *Cephalalgia*. 2015;35:1181-1191.
3. El-Chammas K, Keyes J, Thompson N, Vijayakumar J, Becher D, Jackson JL. Pharmacologic treatment of pediatric headaches: A meta-analysis. *JAMA Pediatr*. 2013;167:250-258.
4. Powers SW, Coffey CS, Chamberlin LA, et al. Trial of amitriptyline, topiramate, and placebo for pediatric migraine. *N Engl J Med*. 2017;376:115-124.
5. Fisher E, Law E, Dudeney J, et al. Psychological therapies for the management of chronic and recurrent pain in children and adolescents. *Cochrane Database Syst Rev*. 2018;9:CD003968.
6. Trautmann E, Lackschewitz H, Kroner-Herwig B. Psychological treatment of recurrent headache in children and adolescents – A meta-analysis. *Cephalalgia*. 2006;26:1411-1426.
7. Stubberud A, Varkey E, McCrory DC, Pedersen SA, Linde M. Biofeedback as prophylaxis for pediatric migraine: A meta-analysis. *Pediatrics*. 2016;138:e20160675.
8. Schwartz MS, Andrasik F. *Biofeedback: A Practitioner's Guide*. New York: Guilford Publications; 2017.
9. Stubberud A, Linde M. Digital technology and mobile health in behavioral migraine therapy: A narrative review. *Curr Pain Headache Rep*. 2018;22:66.
10. Lalloo C, Jibb LA, Rivera J, Agarwal A, Stinson JN. "There'sa pain app for that": Review of patient-targeted smartphone applications for pain management. *Clin J Pain*. 2015;31:557-563.
11. Minen MT, Torous J, Raynowska J, et al. Electronic behavioral interventions for headache: A systematic review. *J Headache Pain*. 2016;17:51.
12. Mosadeghi-Nik M, Askari MS, Fatehi F. Mobile health (mHealth) for headache disorders: A review of the evidence base. *J Telemed Telecare*. 2016;22: 472-477.
13. Stubberud A, Omland PM, Tronvik E, Olsen A, Sand T, Linde M. Wireless surface electromyography and skin temperature sensors for biofeedback treatment of headache: Validation study with stationary control equipment. *JMIR Biomed Eng*. 2018;3:e1.
14. Gillinov S, Etiwy M, Wang R, et al. Variable accuracy of wearable heart rate monitors during aerobic exercise. *Med Sci Sports Exerc*. 2017;49:1697-1703.
15. Wang R, Blackburn G, Desai M, et al. Accuracy of wrist-worn heart rate monitors. *JAMA Cardiol*. 2017;2:104-106.
16. de la Vega R, Miró J. mHealth: A strategic field without a solid scientific soul. A systematic review of pain-related apps. *PLoS ONE*. 2014;9:e101312.
17. Headache Classification Committee of the International Headache Society (IHS). The International Classification of Headache Disorders, 3rd edition. *Cephalalgia*. 2018;38:1-211.
18. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: A new tool for assessing the quality of health mobile apps. *JMIR mHealth uHealth*. 2015;3:e27.
19. Nestoriuc Y, Martin A. Efficacy of biofeedback for migraine: A meta-analysis. *Pain*. 2007;128:111-127.
20. Rains JC. Change mechanisms in EMG biofeedback training: Cognitive changes underlying

improvements in tension headache. *Headache*. 2008; 48:735-736.
21. Fleming S, Thompson M, Stevens R, et al. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: A systematic review of observational studies. *Lancet*. 2011;377:1011-1018.
22. Semizel E, Öztürk B, Bostan OM, Cil E, Ediz B. The effect of age and gender on the electrocardiogram in children. *Cardiol Young*. 2008;18:26-40.
23. Zapata BC, Fernández-Alemán JL, Idri A, Toval A. Empirical studies on usability of mHealth apps: A systematic literature review. *J Med Syst*. 2015;39:1.
24. Zapata BC, Fernández-Alemán JL, Toval A, Idri A. Reusable software usability specifications for mHealth applications. *J Med Syst*. 2018;42:45.
25. Schmidt B. Proof of principle studies. *Epilepsy Res*. 2006;68:48-52.
26. Huguet A, McGrath PJ, Wheaton M, et al. Testing the feasibility and psychometric properties of a mobile diary (myWHI) in adolescents and young adults with headaches. *JMIR mHealth uHealth*. 2015;3:e39.
27. Hughes CM, Hintze A, Padilla A, et al. Development of a mHealth system for post-stroke upper limb rehabilitation in medically underserved populations: An iterative usability study. In: *2018 IEEE Global Humanitarian Technology Conference (GHTC)*: IEEE; 2018:1-8.
28. Alwashmi MF, Hawboldt J, Davis E, Fetters MD. The iterative convergent design for mobile health usability testing: Mixed-methods approach. *JMIR mHealth uHealth*. 2019;7:e11656.
29. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth app usability questionnaire (MAUQ): Development and validation study. *JMIR mHealth uHealth*. 2019;7:e11500.
30. Furnham A. Response bias, social desirability and dissimulation. *Pers Individ Dif*. 1986;7:385-400.
31. Knowles ES, Nathan KT. Acquiescent responding in self-reports: Cognitive style or social concern? *J Res Pers*. 1997;31:293-301.
32. Tfelt-Hansen P, Pascual J, Ramadan N, et al. Guidelines for controlled trials of drugs in migraine: third edition. A guide for investigators. *Cephalalgia*. 2012;32:6-38.
33. Penzien DB, Andrasik F, Freidenberg BM, et al. Guidelines for trials of behavioral treatments for recurrent headache, first edition: American Headache Society Behavioral Clinical Trials Workgroup. *Headache*. 2005;45(Suppl. 2):S110-S132.
34. Schmettow M. Sample size in usability studies. *Commun ACM*. 2012;55:64-70.
35. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. In: *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*: ACM; 1993:206-213
36. Hwang W, Salvendy G. Number of people required for usability evaluation: The 10±2 rule. *Commun ACM*. 2010;53:130-133.
37. Chatzipavlou IA, Christoforidou SA, Vlachopoulou M. A recommended guideline for the development of mHealth apps. *mHealth*. 2016;2:21.
38. Shneiderman B. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. London: Pearson Education India; 2010.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.