Gamification as a tool for enhancing graduate medical education

Christa R Nevin,[1] Andrew O Westfall,[1] J Martin Rodriguez,[1] Donald M Dempsey,[1] Andrea Cherrington,[1] Brita Roy,[1] Mukesh Patel,[2] James H Willig[1]

[1]University of Alabama, Birmingham, USA
[2]University of Maryland, Baltimore, USA

Correspondence to
Dr James H Willig, Department of Medicine, Division of Infectious Diseases, University of Alabama School of Medicine (UAB), BBRB 220B, 1530 3rd Ave S, Birmingham, AL 35294-2050, USA; jwillig@uab.edu

Open Access
Scan to access more
free content

ABSTRACT

Introduction The last decade has seen many changes in graduate medical education training in the USA, most notably the implementation of duty hour standards for residents by the Accreditation Council of Graduate Medical Education. As educators are left to balance more limited time available between patient care and resident education, new methods to augment traditional graduate medical education are needed.

Objectives To assess acceptance and use of a novel gamification-based medical knowledge software among internal medicine residents and to determine retention of information presented to participants by this medical knowledge software.

Methods We designed and developed software using principles of gamification to deliver a web-based medical knowledge competition among internal medicine residents at the University of Alabama (UA) at Birmingham and UA at Huntsville in 2012–2013. Residents participated individually and in teams. Participants accessed daily questions and tracked their online leaderboard competition scores through any internet-enabled device. We completed focus groups to assess participant acceptance and analysed software use, retention of knowledge and factors associated with loss of participants (attrition).

Results *Acceptance:* In focus groups, residents (n=17) reported leaderboards were the most important motivator of participation. *Use:* 16 427 questions were completed: 28.8% on Saturdays/Sundays, 53.1% between 17:00 and 08:00. *Retention of knowledge:* 1046 paired responses (for repeated questions) were collected. Correct responses increased by 11.9% ($p<0.0001$) on retest. Differences per time since question introduction, trainee level and style of play were observed. *Attrition:* In ordinal regression analyses, completing more questions (0.80 per 10% increase; 0.70 to 0.93) decreased, while postgraduate year 3 class (4.25; 1.44 to 12.55) and non-daily play (4.51; 1.50 to 13.58) increased odds of attrition.

Conclusions Our software-enabled, gamification-based educational intervention was well accepted among our millennial learners. Coupling software with gamification and analysis of trainee use and engagement data can be used to develop strategies to augment learning in time-constrained educational settings.

INTRODUCTION

The last decade has seen many changes in graduate medical education training in the USA. Among the most prominent, the Accreditation Council for Graduate Medical Education (ACGME) issued guidelines in July 2003, and again in July 2011, that placed restrictions on the number of hours worked by medical residents during their training.[1] Another important change is the arrival of millennial students to graduate medical education settings. This generation of learners has had ubiquitous access to information technology throughout their education. Studies of the educational impact of ACGME work hour guidelines have been inconclusive and questions persist among educators on how to best prepare millennial residents in this new work hour-regulated educational environment.[2–8] As educators are left to balance the more limited time available between patient care and resident education, new methods to augment traditional graduate medical education are needed to best prepare residents within the new ACGME-mandated environment.

Gamification is the use of elements of game design to increase user engagement. Gamification has been successfully incorporated into medical and scientific endeavours in recent years, from health/ fitness and patient education applications, to genome comparisons (Phylo), protein structure prediction (Foldit) and malaria parasite quantification.[9–12] Due to its proven ability to improve motivation, participation and time investment across multiple settings, we incorporated elements of gamification into the design of software that allowed our residents to participate in a medical knowledge competition with their peers in order to encourage extracurricular learning.[13–15] We used the conceptual frameworks of user-centred design and situational relevance to achieve meaningful gamification, including connecting with users in multiple ways and aligning our 'game' with our residents' backgrounds and interests in furthering their education. The purpose of this study was to assess acceptance and use of a novel gamification-based medical knowledge software designed to supplement traditional graduate medical education among internal medicine (IM) residents and to determine retention of information on subsequent retest.[16 17]

METHODS

Setting

Our study was conducted at two IM training programmes in the USA: the Tinsley Harrison Internal Medicine residency programme at the University of Alabama at Birmingham (UAB) and the University of Alabama at Huntsville (UAH) programme. Inpatient rotations in both general medicine and subspecialties are completed at tertiary care centres, and teams consist of attending physicians, residents (postgraduate year 1 (PGY1)–postgraduate year 3

(PGY3) and medical students. All residents currently completing their IM training (n=128 at UAB and n=24 at UAH) were invited to participate via email or announcements at programme conferences.

Gamification and software design

We named our software Kaizen-Internal Medicine (Kaizen-IM). Kaizen, a Japanese word from the quality improvement literature, signifies the need for continuous daily advancement, a concept analogous to the principle of lifelong learning we seek to inculcate in our residents. Gamification elements included in Kaizen-IM included (1) voluntary participation; (2) explicit, consistent, software-enforced rules of competition for all participants; (3) immediate feedback (response correct or incorrect, followed by explanation of key concepts); (4) team participation with trainees divided into groups as well as individual participation and (5) participants could increase in rank or level (badges granted for score milestones or other achievements). Kaizen-IM could be accessed via the UAB residency website or a link provided in weekly emails after January 2013. Upon registration, participants could input a unique username for display on the leaderboard so that they could remain anonymous. Additionally, they identified their trainee level (PGY1–3).

Questions and competition

Our Kaizen-IM season lasted throughout the 2012–2013 academic year and was divided into three rounds (rounds 1–3). Each round included general IM questions and questions from three medical subspecialties. Questions were written by faculty and emphasised clarity, brevity and were followed by a concise explanation of the correct response (figure 1). Questions were published at 00:01 daily, there was no time limit for response and unanswered questions remained to be addressed at the trainee's convenience. Participants could review prior questions/

Figure 1 Sample of Kaizen-Internal Medicine (Kaizen-IM) graphical user interface. (A) Kaizen-IM question and response. (B) Kaizen-IM leaderboard.

answers and earned badges at any time. Residents competed with one another as individuals and were also divided into six predetermined teams based on faculty advisor for team competition.

Rounds

Round 1 (20 August 2012–13 October 2012): included IM, rheumatology, infectious diseases and pulmonary medicine. A total of 107 questions were delivered at a rate of approximately two per day. Only UAB trainees participated in round 1.

Round 2 (14 January 2013–26 February 2013): included IM, cardiology, gastroenterology and haematology-oncology. At a rate of approximately two per day, 84 questions were delivered. UAH residents were invited to join the competition in round 2.

Round 3 (2 April 2013–13 May 2013): included IM, nephrology, neurology and endocrinology. At a rate of roughly three per day, 117 questions were delivered to UAB and UAH residents. Two questions were new and one was a prior question from either round 1 or 2.

Upon logging into Kaizen-IM, participants viewed a dynamically adjusted leaderboard showing contestants with scores just above (within reach) and immediately below (close enough to pass) their own. Five points were awarded for each correct answer. Extra points and badges were awarded for consistency or daily completion of questions for predetermined intervals (marathons) and for achieving benchmarks of consecutive correct responses (hotstreaks). Scores determined progression through 13 ranks, each accompanied by a new level badge awarded at roughly 100-point increments. Weekly 'status of competition' email notifications were sent to highlight how individuals and teams were faring, seeking to remind and motivate participation. At the end of the academic year, following three rounds of competition, the team with the most cumulative points had their names engraved on a plaque.

Data

Participant-level and question-level data were recorded automatically when Kaizen-IM was used. Participant data included player identification, trainee level (PGY 1–3) and team. Question data included date/time, device used (smartphone, laptop and other devices), badges earned and response accuracy. 'Play styles' were characterised based on the number of questions answered: daily (answered questions within 1–2 days), catch-up (completed 2–6 days of questions) and binge (answered >7 days of questions).

Statistical analyses

Analyses focused on acceptance, use, determination of factors associated with loss of players (attrition) and retention of knowledge. Because traditional tests of normality such as the Kolmogorov–Smirnov test, the Anderson–Darling test and the Shapiro–Wilk test are subject to low power, particularly when the sample size is small, continuous outcome measures were graphically assessed for normality by investigating the distributional form of the outcomes using histograms. When normality assumptions were not met, the appropriate rank-based Wilcoxon test was used.

Acceptance

We invited residents at UAB who had participated (ie, answered ≥1 question) to take part in four focus groups at the conclusion of round 1. Focus groups were conducted in November–December 2012. Each group was limited to no more than eight residents and a $35 gift certificate was given as a participation incentive. Focus groups were audio recorded and transcribed

verbatim. Common themes were then coded by two independent reviewers (CRN and JHW or AC) using a combined deductive/inductive approach.

Use

Medians and the corresponding IQR, 25th to 75th centile, were used to detail Kaizen-IM software use per round (1–3) and trainee level (PGY1–3). Because utilisation data were counts, Poisson regression models were used to test for differences in use by trainee level for each round. A Bonferroni correction was applied to control the overall type 1 error rate across the 27 individual Poisson regression models. p Values $<0.05/27 = 0.002$ were considered statistically significant.

Notification effect: In order to assess the impact of weekly 'status of the competition' emails, we analysed the differences in the total number of questions completed and the number of users participating the day before and the day of week 'status of the competition' emails. These continuous outcome measures (number of questions completed and number of users) were graphically assessed for normality by investigating the distributional form of the outcomes using histograms. Because normality assumptions were not met using this method, Wilcoxon signed-rank tests were used.

Badge effect: In order to test whether earning badges led to greater use of Kaizen-IM, after a specific number of questions had been released in each round, we categorised participants into those who had earned a badge and those who had not earned a badge and then quantified the proportion of remaining questions completed in the round. Differences in the number of questions answered between those who had earned a badge and those who had not were graphically assessed for normality by investigating the distributional form of the outcome using histograms. Because normality assumptions were not met using this method, two-sided Wilcoxon rank-sum tests were used to analyse these differences. In addition, we calculated the Hodges–Lehmann estimate of the difference and associated CI between those who earned a badge and those who did not.

Attrition

Two analyses of factors associated with loss of players or attrition were performed: attrition per debut round and longitudinal attrition.

Attrition per debut round: This analysis included debut players from round 1 (UAB) and round 2 (UAB, UAH) and assessed whether or not they participated in the Kaizen-IM round subsequent to their debut (ie, round 2 if they debuted in round 1 or round 3 if they debuted in round 2). Univariate and multivariable logistic regression models were fit to determine factors associated with player loss after one round of play. Variables included in the multivariable model include player class, debut round, predominant play style, per cent of correct answers and per cent of questions completed.

Longitudinal attrition: Using only players who debuted in round 1, a three-level measure to evaluate overall player attrition was created based on the number of rounds completed after the debut round, zero, one or two. Univariate and multivariable ordinal logistic regression models were fit to determine factors associated with longitudinal attrition.

Retention of knowledge

To test retention of knowledge, one prior question (from rounds 1 or 2) was added to the two new daily questions in round 3. For each player, we computed the per cent of correct answers the first (during round 1 or 2) and second time (during round 3) the

questions were encountered. Assumptions of normality were met based on histograms assessing the distributional form of the outcome and paired t tests were used to test for differences in the first and second responses overall. We also tested for differences stratified by trainee level (PGY1–3) and by the round during which the question was initially answered (round 1 or 2). A Bonferroni correction was applied to control the overall type 1 error rate across the 12 individual paired t tests. p Values <0.05/ 12 = 0.004 were considered statistically significant.

All analyses were completed using SAS software V.9.3.

RESULTS
Acceptance

Seventeen residents who participated in round 1 took part in four focus groups. Among the features singled out as enjoyable about Kaizen-IM, they noted, 'Concise, clinically-relevant questions that could be answered quickly during resident down time' and 'a mixture of text-based and image-based… and of easy and more challenging questions'. Residents also reported receiving immediate feedback on their responses ('why a given answer was correct') aided learning. Residents enjoyed the abilities to review previously answered questions, choose anonymous names to be displayed on the leaderboard and answer questions at a date of their convenience.

The single most important motivator for participation identified was the leaderboard. Residents enjoyed the opportunity to compete with their peers as both individuals and teams. While none reported that competition alone was de-motivating, those that found themselves or their teams towards the bottom of the leaderboard reported a loss of motivation as the competition progressed. Residents reported that earning new rank/level badges was less motivating than the leaderboard. Several noted initial excitement at earning badges, but lost interest over time. Reasons for this included lack of understanding of what was required to earn badges and how much progress had been made towards earning a badge.

The primary difficulty reported and a common cause for declining participation by round 1 players was Kaizen-IM accessibility. At the onset of round 1, residents received an email including the link to the Kaizen-IM website and a link was posted on the UAB residency website for easy accessibility. Many individuals reported they did not visit the residency website regularly necessitating repeated searches of email inboxes which discouraged participation.

Table 1 Kaizen-Internal Medicine use stratified by postgraduate year (PGY1–3) of training and round of the competition (1–3)

	Total	PGY1	PGY2	PGY3	p Value*
Round 1	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	
107 questions	**(N=92)**	**(N=36)**	**(N=33)**	**(N=23)**	
Questions answered†	97.0 (42.0 to 107.0)	97.5 (54.5 to 107.0)	97.0 (23.0 to 107.0)	87.0 (57.0 to 107.0)	0.443
Per cent correct	71.8 (61.7 to 78.9)	70.2 (58.7 to 76.7)	67.3 (61.0 to 75.6)	78.7 (73.8 to 86.0)	0.001
Days to answer	4.3 (1.5 to 8.3)	3.1 (1.1 to 8.2)	4.0 (1.7 to 6.1)	6.4 (3.1 to 12.2)	0.810
Play duration‡	42.0 (19.5 to 50.0)	43.5 (26.0 to 50.0)	39.0 (8.0 to 50.0)	41.0 (30.0 to 51.0)	0.210
Days between sessions‡	3.1 (1.7 to 7.0)	2.2 (1.5 to 3.8)	4.0 (2.5 to 7.0)	6.0 (3.0 to 10.0)	<0.001
Questions per day†	7.2 (4.0 to 11.9)	4.6 (3.2 to 8.0)	8.0 (5.4 to 13.4)	8.8 (3.8 to 13.7)	0.083
≥1 question answered‡	9.0 (4.0 to 20.5)	15.0 (6.5 to 26.5)	7.0 (3.0 to 15.0)	8.0 (4.0 to 12.0)	0.080
Badges earned†	2.0 (1.0 to 3.0)	2.0 (1.5 to 4.0)	2.0 (0.0 to 3.0)	3.0 (2.0 to 4.0)	0.105
Devices used†	2.0 (1.0 to 3.0)	2 .0 (1.0 to 3.0)	2.0 (1.0 to 3.0)	2.0 (1.0 to 3.0)	0.811
Round 2§	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	
84 questions	**(N=90)**	**(N=42)**	**(N=32)**	**(N=16)**	
Questions answered†	65.5 (36.0 to 84.0)	57.0 (36.0 to 84.0)	69.0 (29.5 to 84.0)	83.0 (54.0 to 84.0)	0.482
Per cent correct	66.7 (57.7 to 73.8)	68.6 (65.0 to 74.2)	63.1 (54.4 to 66.7)	66.1 (58.9 to 82.1)	0.025
Days to answer	5.8 (1.7 to 11.4)	5.9 (1.7 to 13.5)	6.1 (1.8 to 14.6)	3.8 (1.0 to 8.1)	0.429
Play duration‡	34.0 (18.0 to 40.0)	35.0 (18.0 to 40.0)	32.0 (15.0 to 39.5)	33.0 (23.5 to 40.5)	0.752
Days between sessions‡	3.7 (1.8 to 7.0)	4.4 (2.2 to 7.8)	3.7 (1.7 to 7.1)	3.0 (1.8 to 4.4)	0.281
Questions per session*	6.5 (3.7 to 10.5)	7.6 (4.0 to 11.3)	6.1 (3.3 to 9.8)	5.6 (3.7 to 10.1)	0.564
≥1 question answered‡	8.5 (3.0 to 18.0)	7.5 (3.0 to 18.0)	8.0 (2.5 to 17.0)	10.5 (5.5 to 20.0)	0.919
Badges earned†	2.0 (1.0 to 2.0)	1.0 (1.0 to 2.0)	1.0 (0.0 to 2.0)	2.0 (1.5 to 3.0)	0.124
Devices used†	2.0 (1.0 to 2.0)	2.0 (1.0 to 2.0)	1.5 (1.0 to 3.0)	1.5 (1.0 to 2.5)	0.873
Round 3	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	
117 questions	**(N=55)**	**(N=25)**	**(N=22)**	**(N=8)**	
Questions answered†	108.0 (31.0 to 117.0)	117.0 (72.0 to 117.0)	58.0 (8.0 to 117.0)	73.0 (38.0 to 114.0)	0.049
Per cent correct	60.3 (50.4 to 71.4)	62.9 (54.2 to 72.7)	55.0 (50.0 to 71.4)	62.3 (57.0 to 65.6)	0.471
Days to answer	5.0 (1.2 to 17.2)	4.5 (1.1 to 17.2)	7.7 (2.3 to 18.9)	2.3 (0.5 to 10.2)	0.335
Play duration‡	30.0 (5.0 to 38.0)	34.0 (18.0 to 37.0)	16.0 (4.0 to 33.0)	35.5 (3.5 to 40.0)	0.187
Days between sessions‡	2.6 (1.5 to 4.7)	2.4 (1.6 to 4.4)	3.3 (1.6 to 5.7)	2.2 (1.3 to 3.8)	0.589
Questions per day†	7.3 (4.4 to 13.3)	10.3 (5.1 to 14.6)	6.7 (3.5 to 9.8)	7.5 (4.5 to 14.7)	0.451
≥1 question answered‡	6.0 (2.0 to 16.0)	8.0 (3.0 to 17.0)	4.0 (2.0 to 13.0)	7.0 (2.5 to 20.5)	0.544
Badges earned†	2.0 (0.0 to 4.0)	3.0 (2.0 to 4.0)	1.0 (0.0 to 2.0)	1.5 (0.5 to 2.5)	0.037
Devices used†	1.0 (1.0 to 2.0)	1.0 (1.0 to 2.0)	1.0 (1.0 to 2.0)	1.0 (1.0 to 1.5)	0.402

*Type 3 p values from separate Poisson regression models modelling the count by trainee level.
†Per participant.
‡Unit is days.
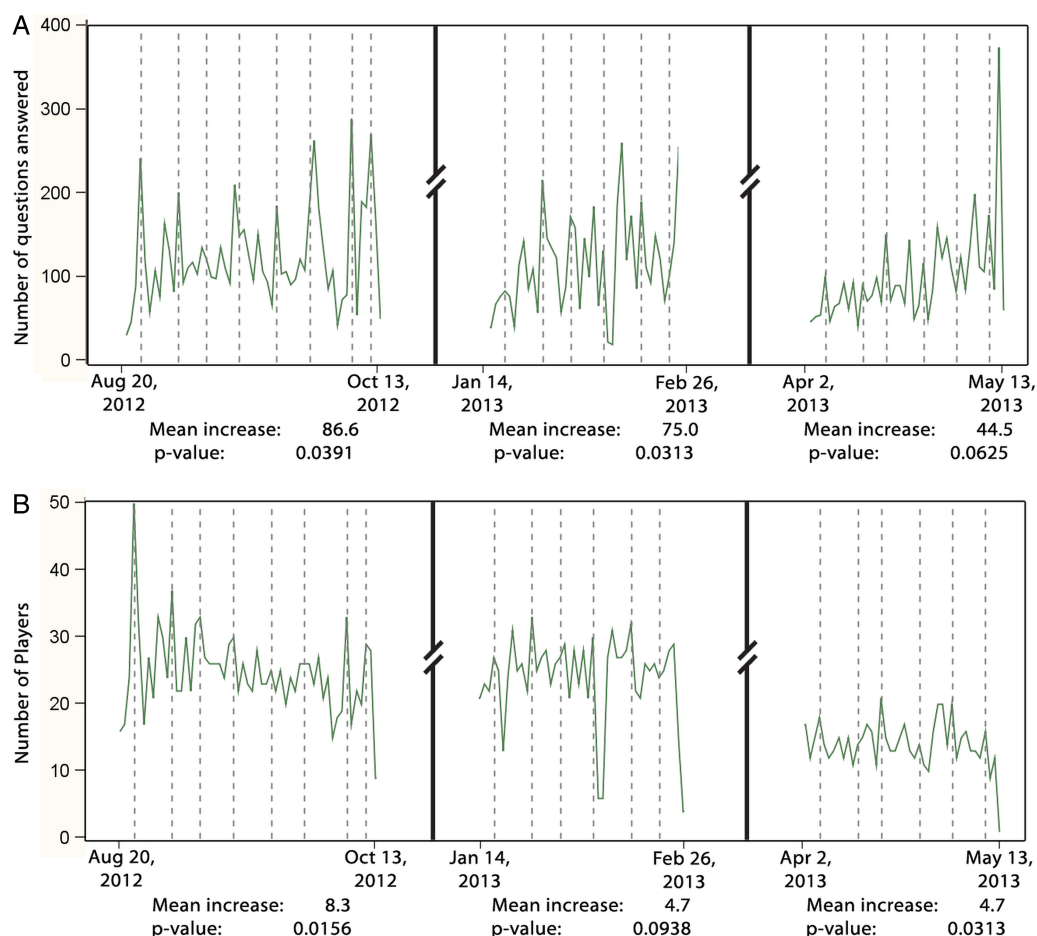§A total of 26 new players were added in round 2.

Figure 2 Effect of weekly 'status of competition' notification emails on the total number of questions answered (A) and total number of players (B) in the day prior to versus the day of each notification email. Each dotted vertical line corresponds to the date a 'status of competition' notification email was sent, while the horizontal line shows resulting fluctuations in number of questions answered (A) and number of players answering questions (B). Wilcoxon signed-rank tests were used to test for statistically significant changes in number of questions completed and user participation in the day prior and the day of the weekly notification emails.

Use

The number of participants declined in each successive round with 92 participants in round 1, 90 in round 2 and 55 in round 3 that delivered 107, 84 and 117 questions, respectively. Participants answered 16 427 questions throughout the season. The percentage of questions completed (questions answered/total players × total questions) was 71 in round 1 (6985/9844), 69 in round 2 (5180/7560) and 66 in round 3 (4262/6435).

Table 1 details use of Kaizen-IM by round (1–3) for each PGY level (1–3) and includes p values from Poisson regression models modelling the count by the trainee level. In round 1, there were statistically significant differences in the number of days between sessions with PGY1 accessing Kaizen-IM most frequently, median (IQR) of 2.2 days (1.5 to 3.8), and PGY3 residents performing best, median (IQR) of 78.7% correct (73.8 to 86). Across rounds, 28.8% of questions were done on Saturdays and Sundays while 53.1% were completed between 17:00 and 08:00.

Notification effect: Wilcoxon signed-rank tests showed statistically significant increases in total questions answered for rounds 1 and 2 and number of participants for rounds 1 and 3, when comparing totals for days before and days when weekly 'status of competition' emails were sent (figure 2).

Badge effect: Wilcoxon rank-sum tests evaluated differences in the proportion of questions completed after earning a badge at a prespecified point in each round. In round 1, after 25 questions, there was no statistically significant difference in the proportion of

subsequent questions completed for those who had earned at least one badge (n=38; median (IQR) of 98.2 (51.2 to 100.0)) versus those who had not earned at least one badge (n=40, median (IQR) of 90.9 (51.2 to 100.0), p=0.36, Hodges–Lehmann estimate of shift=0.0 (0.0 to 7.3)). In round 2, after 25 questions, those who had earned badges (n=12, median (IQR) of 100.0 (100.0 to 100.0)) compared with those who had not (n=64, median (IQR) of 77.1 (32.2 to 100.0)) were found to be more likely to answer a greater proportion of remaining questions (p=0.02, Hodges–Lehmann estimate of shift=5.9 (0.0 to 39.0)). In round 3, after 35 questions, there was no statistically significant difference in the proportion of remaining questions completed (p=0.15, Hodges–Lehmann estimate of shift=0.0 (0.0 to 25.6)) between those who had (n=19, median (IQR) of 100.0 (89.0 to 100.0)) and had not (n=22, median (IQR) of 100.0 (34.1 to 100.0)) earned badges.

Attrition (player loss)

Attrition per debut round: 33% of players did not participate in the round following their debut. In logistic regression models for factors associated with participant loss, we found for every 10% increase in questions completed the odds of attrition decreased (0.81; 0.69 to 0.95). UAB participants who skipped round 1 and debuted in round 2 were at high risk for stopping play (12.69; 1.34 to 120.15) (table 2).

Longitudinal attrition: Analysis of player attrition throughout the season (only UAB players who debuted in round 1) showed

Table 2 Logistic regression analysis of factors associated with attrition (player loss) after the debut round of play among University of Alabama at Birmingham (UAB) and University of Alabama at Huntsville (UAH) residents (n=117) participating in Kaizen-Internal Medicine during the 2012–2013 academic year

	Number of players	Played next round N (%) or median (IQR)	Did not play next round N (%) or median (IQR)	Univariate*	Multivariable†
Player class					
PGY1	48	37 (77%)	11 (23%)	1.0	1.0
PGY2	43	27 (63%)	16 (37%)	1.99 (0.80 to 4.97)	1.54 (0.49 to 4.79)
PGY3	26	14 (54%)	12 (46%)	2.88 (1.04 to 8.02)	2.17 (0.59 to 7.94)
Debut round					
UAB round 1	92	65 (71%)	27 (29%)	1.0	1.0
UAB round 2‡	10	1 (10%)	9 (90%)	21.65 (2.62 to 179.25)	12.69 (1.34 to 120.15)
UAH round 2‡	15	12 (80%)	3 (20%)	0.60 (0.16 to 2.30)	1.33 (0.29 to 6.15)
Predominant play style§					
Daily	46	37 (80%)	9 (20%)	1.0	1.0
Other	71	41 (58%)	30 (42%)	3.01 (1.26 to 7.16)	2.82 (0.86 to 9.28)
Per cent correct**	117	68.5 (58.9 to 78.1)	76.2 (66.7 to 86.0)	1.39 (1.04 to 1.86)	1.31 (0.84 to 2.05)
Per cent completed**	117	98.6 (66.4 to 100.0)	42.1 (10.3 to 75.7)	0.74 (0.65 to 0.84)	0.81 (0.69 to 0.95)

*Results in the 'univariate' column are based on five separate logistic regression models each modelling attrition (Y/N) in the round after debut. p Values for the overall type 3 tests for the three-level categorical variables player class and debut round were 0.11 and 0.08, respectively.
†Results in the 'multivariable' column are based on a single logistic regression model modelling attrition (Y/N) in the round after debut. The model included player class, debut round, predominant play style, per cent correct and per cent completed. p Values for the overall type 3 tests for the three-level categorical variables player class and debut round were 0.49 and 0.08, respectively.
‡ UAB round 2 debut players include individuals who did not play in round 1, but joined in round 2. UAH players were invited to participate for rounds 2 and 3, but did not play in round 1.
§Predominant play style was defined as the play style used by each participant the majority of the time. Daily players most often completed their questions on the day of release; catch-up (2–6 days of questions at a time) and binge (>7 days of questions at a time) are collapsed into the 'other' category.
**Per cent correct and per cent completed included in models as continuous variables. ORs represent per 10% increase.
PGY, postgraduate year.

a loss of 29% (n=27) after round 1 and 28% (n=26) after round 2. In ordinal logistic regression analysis, completing more questions (0.80 per 10% increase; 0.70 to 0.93) was associated with lower likelihood of attrition. PGY3 versus PGY1 class (4.25; 1.44 to 12.55) and non-daily play (4.51; 1.50 to 13.58) were associated with increased odds of player loss. The proportional odds assumption was met for all ordinal models.

Retention of knowledge

Overall, 50 participants answered ≥1 retention question both times it was presented for a total of 1046 paired responses.

On average, players' answers were correct 11.9% (p<0.001) more frequently when a question was reintroduced. Benefits were seen for both reintroduced round 1 questions (10.2% increase, p=<0.001) and round 2 questions (17.4% increase, p<0.001) (table 3).

DISCUSSION

In a time of work hour restrictions where resident hospital presence is strictly regulated, we sought to capitalise on the technological savvy of millennial learners with a new approach to supplement traditional medical instruction.[2] We developed web-

Table 3 Overall and round-specific retention of knowledge among participants who answered the questions initially (rounds 1 or 2) and on retest (round 3)

	Number of players (answer pairs)	Initial per cent correct Mean (SD)	Subsequent per cent correct Mean (SD)	Mean change (95% CI)	p Value*
Overall	50 (1046)	40.5 (20.1)	52.3 (23.1)	11.9 (6.7 to 17.0)	<0.001
PGY1	25 (567)	49.6 (18.5)	59 (21.6)	9.5 (2.3 to 16.7)	0.012
PGY2	18 (309)	26 (14.7)	41.8 (22.4)	15.8 (6.4 to 25.2)	0.002
PGY3	7 (170)	45.2 (17.7)	55.4 (23.0)	10.2 (−7.0 to 27.3)	0.197
Round 1; 20 August 2012–13 October 2012					
Overall	39 (546)	37.8 (18.1)	47.9 (21.9)	10.2 (4.5 to 15.8)	<0.001
PGY1	18 (271)	44.1 (15.3)	50 (20.2)	5.9 (−1.8 to 13.6)	0.123
PGY2	14 (165)	25.8 (16.3)	40.9 (23.4)	15.1 (4.1 to 26.1)	0.011
PGY3	7 (110)	45.5 (17.8)	56.7 (21.6)	11.2 (−6.4 to 28.8)	0.170
Round 2; 14 January 2013–26 February 2013					
Overall	32 (500)	46.2 (19.5)	63.6 (22.0)	17.4 (11.1 to 23.7)	<0.001
PGY1	19 (296)	52.2 (18.7)	69.3 (20.3)	17.1 (7.4 to 26.8)	0.002
PGY2	9 (144)	32.9 (14.8)	52.8 (20.4)	19.9 (8.4 to 31.3)	0.004
PGY3	4 (60)	47.9 (22.4)	61.3 (29.6)	13.4 (−1.3 to 28.0)	0.062

*Paired t test.
PGY, postgraduate year.

based medical knowledge competition software (Kaizen-IM) and incorporated gamification strategies into our design in an effort to encourage resident participation. Each of the 16 427 questions completed and the ensuing explanation of the key point represented additional teaching interactions that otherwise would not have occurred in the academic year. With over a quarter of questions answered on weekends and over half between 17:00 and 08:00, it is clear we were able to engage residents outside of typical programme didactics. We found both qualitative and quantitative evidence of a beneficial effect of gamification strategies on Kaizen-IM use. Taken together, these data suggest that our gamification-based educational software effectively engaged residents, similar to the effects of gamification observed in other settings.[9 18 19] In addition, analyses of Kaizen-IM data provided insights into modifiable factors associated with learner attrition and retention of knowledge that may serve to further enhance the educational benefits of this strategy for our trainees.

While gamification has been used in educational settings, reports of its integration and effect in medical education settings are limited.[18 19] In focus groups, residents highlighted the leaderboard as highly motivational, and the ability to compete as both individuals and teams allowed Kaizen-IM to appeal to a greater number of users. Of note, displaying each participant's team affiliation on the leaderboard may have led to increased investment through promoting accountability between team members.[10] We saw an effect of earning badges and of our weekly 'status of competition' notification emails on use. While those who had earned badges at the time of analyses completed greater numbers of questions, a statistically significant difference was only seen in round 2. Weekly 'status of competition' notifications were associated with statistically significant increases or trends towards both an increased number of questions completed and competitors playing the day after each communiqué. This suggests that regular notifications have value in enhancing participant retention in software-based educational initiatives, though further research is needed to delineate the most effective time interval for contacting trainees.

We designed our analyses of Kaizen-IM data using insights derived from observing the highly competitive 'Free to play' (F2P) software space. F2P games are ubiquitous and allow users access to software at no cost, while developers meticulously analyse player data for insights on how to increase user acquisition and engagement.[20 21] User data are then used to tweak software parameters to further engage users and maximise playing time. Analysis of Kaizen-IM data can be viewed in a similar light as it provides insight into how to initially and longitudinally engage residents to encourage repeated investments of time-answering questions. The data captured through the use of technology-driven learning approaches such as Kaizen-IM coupled with ongoing analyses represent new opportunities for educators to monitor and assess learner engagement and performance. Strategies such as Kaizen-IM will be able to supplement resident assessment beyond the current periodic evaluations offered by yearly in-service and eventually board certification examinations. In the future, such approaches may provide an early indication of specific educational needs in focused subject areas, allowing timely targeted learning that will benefit residents. We posit that an ongoing technology-based strategy to identify deficiencies and provide ongoing targeted supplementary education may ultimately lead to better-trained providers able to achieve improved evidence-based care and outcomes for their patients.

The statistically significant increases in correct responses upon question retesting suggest participant knowledge retention. Consistent with previous reports of decreased information recall over time in healthcare education settings (decay of knowledge), overall improvement in scores was higher for more recent (round 2; 17.4%) versus older questions (round 1; 10.2%).[22 23] Resident-level comparisons generally found less statistically significant improvements in the PGY3 than other classes. This may be due to sample size, but other hypotheses such as PGY3 having decreased engagement due to competing priorities (associated with the proximity of graduation and preparation for next steps) or greater Kaizen-IM effectiveness earlier in training must be considered. Further exploration of factors affecting retention of knowledge in technology-driven educational initiatives will lead to recognition of how best and often to deploy content to maximise long-term recall of information among our technology-savvy millennial learners.[2 24 25]

The limitations of our study include the restriction of analyses to two residency sites both in the southeastern USA and encompassing a single medical specialty (IM). However, our software is a platform where educational content delivered to residents can be changed to accommodate other graduate medical education disciplines that are in the process of adapting to new work hour requirements. Residents participated in Kaizen-IM voluntarily and thus our sample size was limited only to those residents that chose to participate. Additionally, our data encompass only one academic year of Kaizen-IM use. It is possible that a larger sample size of participants and/or a longer period of data collection could affect results. While it is certainly possible we have failed to identify effects that really exist (type II errors), we feel we have shown that residents are not only open to supplemental educational tools such as Kaizen-IM but that they do use them and retain some of the knowledge. In assessing use of Kaizen-IM, we were able to identify the impact of weekly 'status of the competition' emails on participation but we cannot account for the impact of internal communications among team members. Finally, regarding analyses of player attrition and retention of knowledge, we can only describe associations between variables and cannot denote causality. Our study adds to the literature by reporting on a novel gamification-based, software-driven instructional strategy and identifies methods by which to analyse and interpret participant data in order to glean insight into how to maximise learning with such techniques. Additional studies are needed to better understand the impact of Kaizen-IM on more objective educational measures such as board exam scores. Further study into the incorporation of elements of other conceptual models beyond user-centred design, situational relevance and motivation models is also needed to best determine how to enhance learning through gamification-based interventions.

CONCLUSION

Our Kaizen-IM software successfully incorporated elements of gamification and engaged a large number of residents in a medical knowledge competition facilitating acquisition of new knowledge, often outside of regular work/teaching hours. Such educational software platforms, potentially coupled with a metrics-driven approach to analysing resident utilisation data, can be used to deliver and reinforce critical concepts while providing educators with new tools to augment traditional medical education. Such innovations may aid in the education of medical residents whose in-hospital hours are more limited than those of prior generations and support the continued training of the highest quality healthcare providers.

Main messages

▶ With resident work hour regulations restricting time in training hospitals in the USA, new educational approaches are needed to supplement traditional medical instruction.

▶ We used principles of gamification (the use of game design principles to increase engagement) to design internet accessible educational software (Kaizen-IM) engaging residents in medical knowledge competition.

▶ With over a quarter of questions answered on weekends and half between (17:00 and 08:00) we were able to engage residents outside program didactics, found both qualitative and quantitative evidence of a beneficial effect of gamification strategies on use and found statistically significant increases in knowledge retention over time.

▶ Such educational software platforms, coupled with metrics driven approaches to analysing resident utilisation data provide new tools to deliver, reinforce and augment traditional medical education.

Current research questions

▶ Do gamification-based educational interventions designed to supplement traditional medical education improve in-service and/or board examination scores when used?

▶ What level of participation in gamification-based education interventions is required to improve in-service and/or board examination scores?

▶ What elements of gamification are most effective in encouraging long-term use of gamification-based educational interventions by medical residents?

Key references

▶ Eckleberry-Hunt J, Tucciarone J. The challenges and opportunities of teaching "Generation Y". *J Grad Med Educ* 2011;3:458–61.

▶ Nevin CR, Cherrington A, Roy B, *et al*. A qualitative assessment of internal medicine resident perceptions of graduate medical education following implementation of the 2011 ACGME duty hour standards. *BMC Med Educ* 2014;14:84.

▶ Schrope M. Solving tough problems with games. *Proc Natl Acad Sci USA* 2013;110:7104–6.

▶ Lovell N. *Your quick guide to metrics*. Games brief, 4 Jun 2012. http://www.gamesbrief.com/2012/06/ your-quick-guide-to-metrics/ (accessed 29 May 2013).

▶ Mohr NM, Moreno-Walton L, Mills AM, *et al*; Society for Academic Emergency Medicine A, Generational Issues in Academic Emergency Medicine Task F. Generational influences in academic emergency medicine: teaching and learning, mentoring, and technology (part I). *Acad Emerg Med* 2011; 18:190–9.

Correction notice This article has been corrected since it was published Online First. The correct figure 1B has now been included.

REFERENCES

1 The members of the ACGME Task Force on Quality Care and Professionalism. *The ACGME 2011 Duty Hour Standard: Enhancing Quality of Care, Supervision and Resident Professional Development*. Chicago: ACGME, 2011. http://www.acgme.org/acgmeweb/Portals/0/PDFs/jgme-monograph[1].pdf

2 Eckleberry-Hunt J, Tucciarone J. The challenges and opportunities of teaching "Generation Y". *J Grad Med Educ* 2011;3:458–61.

3 Fitzgibbons SC, Chen J, Jagsi R, *et al*. Long-term follow-up on the educational impact of ACGME duty hour limits: a pre-post survey study. *Ann Surg* 2012;256:1108–12.

4 Fletcher KE, Underwood W IIIrd, *et al* Effects of work hour reduction on residents' lives: a systematic review. *JAMA* 2005;294:1088–100.

5 Goitein L, Shanafelt TD, Wipf JE, *et al*. The effects of work-hour limitations on resident well-being, patient care, and education in an internal medicine residency program. *Arch Intern Med* 2005;165:2601–6.

6 Lin GA, Beck DC, Stewart AL, *et al*. Resident perceptions of the impact of work hour limitations. *J Gen Intern Med* 2007;22:969–75.

7 Vaughn DM, Stout CL, McCampbell BL, *et al*. Three-year results of mandated work hour restrictions: attending and resident perspectives and effects in a community hospital. *Am Surg* 2008;74:542–6; discussion 546–547.

8 Nevin CR, Cherrington A, Roy B, *et al*. A qualitative assessment of internal medicine resident perceptions of graduate medical education following implementation of the 2011 ACGME duty hour standards. *BMC Med Educ* 2014;14:84.

9 Schrope M. Solving tough problems with games. *Proc Natl Acad Sci USA* 2013;110:7104–6.

10 Gamification of Education. http://gamification.org/education/ (accessed 29 May 2013).

11 Shane K. *Three more ways gamification in healthcare is making a difference*. Gamification Co, 20 Mar 2013. http://www.gamification.co/2013/03/20/gamification-in-healthcare/ (accessed 29 May 2013).

12 Gopaladesikan S. *A comprehensive overview of how games help healthcare in 2013*. Gamification Co, 14 Mar 2013. http://www.gamification.co/2013/03/14/how-games-help-healthcare/ (accessed 29 May 2013).

13 Gamification: How effective is it? http://www.slideshare.net/ervler/gamification-how-effective-is-it (accessed 29 May 2013).

14 Design in health gamification: motivating behavior change. 16 Aug 2012. http://ayogo.com/blog/design-in-health-gamification-motivating-behaviour-change/ (accessed 29 May 2013).

15 Steer D. Effective gamification improves learning. 1 Oct 2013. http://prezi.com/pa4428w3mnff/effective-gamification-improves-learning/ (acessed 29 May 2013).

16 Bordage G. Conceptual frameworks to illuminate and magnify. *Med Educ* 2009;43:312–19.

17 Nicholson S. A user-centered theoretical framework for meaningful gamification. Presented at *Games+Learning+Society 8.0*, Madison, WI. Jun 2012. http://scottnicholson.com/pubs/meaningfulframework.pdf

18 Odenweller CM, Hsu CT, DiCarlo SE. Educational card games for understanding gastrointestinal physiology. *Am J Physiol* 1998;275(6 Pt 2):S78–84.

19 Snyder E, Hartig JR. Gamification of board review: a residency curricular innovation. *Med Educ* 2013;47:524–5.

20 Lovell N. *Your quick guide to metrics*. Games brief, 4 Jun 2012. http://www.gamesbrief.com/2012/06/your-quick-guide-to-metrics/ (accessed 29 May 2013).

21 Bergel S. *Designing for monetization: How to apply THE key metric In social gaming*. PlayNow [blog], 26 Oct 2009. http://shantibergel.com/post/223918739/designing-for-monetization-how-to-apply-the-key-metric (acessed 29 May 2013).

22 Magura S, Miller MG, Michael T, *et al*. Novel electronic refreshers for cardiopulmonary resuscitation: a randomized controlled trial. *BMC Emerg Med* 2012;12:18.

23 Su E, Schmidt TA, Mann NC, *et al*. A randomized controlled trial to assess decay in acquired knowledge among paramedics completing a pediatric resuscitation course. *Acad Emerg Med* 2000;7:779–86.

24 Mohr NM, Moreno-Walton L, Mills AM, *et al*; Society for Academic Emergency Medicine A, Generational Issues in Academic Emergency Medicine Task F. Generational influences in academic emergency medicine: teaching and learning, mentoring, and technology (part I). *Acad Emerg Med* 2011;18:190–9.

25 White G, Kiegaldie D. Gen Y learners: just how concerned should we be? *Clin Teach* 2011;8:263–6.