The evaluation of health-oriented serious games and apps: A differentiated approach

Management Summary

This report explores the issue of *differentiated validation of health-oriented serious games and apps* by reviewing relevant scientific literature and existing validation tools and initiatives, and by sharing the results of a small survey targeted to industry and academic professionals with relevant experience.

The report builds upon the position paper by Renger, Veltkamp and Schouten (2015). The position paper was a call from the industry and academic sector for further research into the development of a more nuanced and up-to-date methodology for risk assessment and validation, one that follows the different types of games and apps that are now emerging. This report focuses specifically on *validation*, i.e., the study of whether claims about the efficacy of the designed game or app can be empirically substantiated.

Over the past decade close to 20 meta-reviews have been published by the scientific community, looking into the effectiveness of serious games in general and health-oriented games and apps more particularly. The overall picture of the effects of serious games, also in the area of health, is positive. Still, more validated measuring instruments/tools are needed, more research is needed on the specific game features to determine their effectiveness, and more randomized controlled trials (RCTs) should be carried out. These and other concerns should be taken into account when tackling differentiated validation.

The aforementioned meta-reviews cover numerous health-oriented serious games and apps, which can be grouped into several goal/intervention categories. Concerning medical professionals or even the general public, there are games/apps aimed at *learning* (obtaining information, gaining knowledge) and *training* (developing and practicing competencies). Concerning patients, there are games/apps aimed at *prevention, diagnosis, treatment/cure* and *care*. Except for diagnosis, all of these goals/interventions can include learning and/or training.

To reach such goals/interventions, four main dimensions need to be taken into consideration to validate any game or app:

- *Usability*, i.e., functionality and accessibility, which is the result of the intended usage, presentation style and technological back-end;
- *Playability*, i.e., whether the end-users have the freedom to express themselves or act as they please within certain boundaries without serious consequences.

- *Efficacy*, i.e., the ability of a game or app to shape attitudes, knowledge, skills and/or behavior of its users towards an intended state.
- *Side-effects*, i.e., all explicitly or implicitly unintended effects a program has on its users.

To further differentiate validation, we make a distinction between two different types of games/apps and their associated validity types:

- Games/apps as *measurement instruments* for providing information about assessment of acquired knowledge, skills or attitude or of a certain condition. For these we recognize the relevance of these validity types:
    - *Content validity*, focusing on the completeness and correctness of the content included in the game/app;
    - *Face validity*, focusing on the strategy that the game/app proposes for attaining its goal at face value;
    - *Construct validity*, focusing on the chosen way of assessing the goal itself;
    - *Concurrent validity*, focusing on the game's/app's design in reaching its goal compared to other comparable, proven methods;
    - *Predictive validity*, focusing on the game's/app's design reaching its goal in multiple, different settings and situations, including outside of the game/app itself.
- Games/apps as *treatment* of, or *therapy or intervention* for a certain condition. For these we recognize the relevance of these validity types:
    - *Statistical conclusion validity*, focusing on the statistically significant relationship between the treatment and the outcome;
    - *Internal validity*, focusing on the causality of the relationship, and that the relationship is not a result of a variable that has not been measured or that we have no control over;
    - *Construct validity*, focusing on whether the treatment reflects the construct of the cause and whether the outcome reflects the construct of the effect;
    - *External validity*, focusing on generalizability outside the scope of the study.

We then also review distinctions between different experimental study designs, particularly the time-honored RCT and eight alternative designs. We also explain in which situations it is possible or prudent to refrain from using RCTs.

Several validation tools and initiatives for health-oriented games/apps have been released or published. We review four tools/initiatives for heuristic evaluation of usability and playability. We subsequently review two Dutch and several British/US government-led (self-)assessment tools/initiatives. We also review three international academic (self-)assessment tools/initiatives.

Most of all these tools/initiatives are not achieving major results so far. Moreover, empirical, experimental studies are not yet supported by these tools/initiatives. New tools/initiatives for the design or support of experimental studies would thus complement existing ones nicely. Experimental studies and the tools/initiatives to make them happen should not take place purely within the confines of a university or other academic environment.

Our main conclusion is that the claims that are attached to a certain game/app determine the type of validity that should be checked, and at the same time the research design that is needed to examine those claims. This leads to a differentiated approach:

- The first question (or claim) is **to check whether the game/app is merely aimed at assessment of the game/app as an instrument**. If so, it would suffice to pay attention to content, construct and concurrent validity of the assessment.
- **The second question (or claim) concerns whether the game/app as intervention is effective or not**. The internal validity has to be demonstrated through a form of experiment, e.g. by using a RCT or next-best alternative design.
- The third question (or claim) is more detailed: here the question is not **simply whether the treatment/intervention is effective, but also an additional claim is examined: what is the effective component of the game?** Is it the gameplay, or is it simply the case that the monitoring activities are responsible for the results, irrespective of the game, etc.

Finally, the Appendix shares results of a small survey of 28 researchers, designers and other employees experienced or concerned with validation in this domain. The majority says that the current ways in which these applications are validated are too expensive and take up too much time. Many see a need for differentiated methods of validation. Many feel that e-health applications that do not broach sensitive topics should be subject to less stringent validating tests than those that do. The majority calls for an exploration of other methods of validation that do not conform to the RCT model.

Table of Contents

1 Introduction 5

2 Reviewing relevant validation literature 7
2.1 Game/app effectiveness 7
2.2 Concerns 11

3 A conceptual framework for health-oriented game/app validation 12
3.1 A goal/intervention categorization 12
3.2 Relevant outcome variables 13
3.3 Relevant validity types 14
3.4 RCTs and other study designs 17

4 Tools for validation 22
4.1 Heuristic evaluation instruments for usability and playability 22
4.2 (Self)Assessment initiatives 23
4.3 Considerations for new validation tools and initiatives 27

5 Conclusion: A differentiated approach 29
5.1 A summary of the findings 29
5.2 Future research 31

References 33

Appendix: Growing Games Validation Survey 2016 38

1 Introduction

How can we best assess the risks and validity of different types of health-oriented serious games and apps before we can bring them to market? This was the main question posed by Renger, Veltkamp and Schouten in their 2015 position paper. The position paper was a call from the industry and academic sector for further research into the development of a more nuanced and up-to-date methodology for risk assessment and validation, one that follows the different types of games and apps that are now emerging. The authors posited that not all health-oriented serious games and apps should require randomized controlled trials (RCTs), an elaborate and costly form of a full experiment. The exact requirements for risk assessment and validation should depend on:

a. The *application domain*: prevention, community care, low or high complexity care, treatment/therapy;
b. The *target audience*: consumers, caregivers, health professionals;
c. The *type of objective or aim* and the *intervention level*: providing information, raising awareness, (self-) diagnosis/assessment, monitoring, therapy, treatment, learning/training;
d. The individual and social *patient status*: healthy, at high risk of a condition, acute or chronic condition, in close contact with partner, relatives, friends, caregivers,
e. If relevant, the *role of the health professional*: actively involved in (part of) the treatment/care, responsible for defining (part of) the treatment/care, responsible for coordinating;
f. The *data profile*: what personal data is stored, where is it sent and stored;
g. The *scope of application*: narrow (the game/app has a narrow focus), broad (the game/app has a wide range of applications, target audiences, objectives, etc.).

The arguments raised in the position paper call for further study, specifically for further review of relevant scientific literature. The question remains what the scientific medical/health and game design communities have already contributed to this issue of differentiated risk assessment and validation: (1) What knowledge can help us further understand or scrutinize the issue?

The position paper also calls for further scientific exploration of whether existing initiatives and tools aid in differentiated risk assessment and validation. The question remains how we can offer a starting point towards a solution: (2) What scientifically sound conceptual framework for differentiated risk assessment and validation can be posed and how can existing, relevant tools for be subsequently valued and complemented?

This report offers preliminary answers to the above two calls. We provide possible directions towards a solution to the issues raised by Renger, Veltkamp and Schouten. We do this by sharing results of our reviews of relevant scientific literature as well as relevant existing tools and initiatives. In the Appendix we also offer the results of a small survey of industry and academic professionals with relevant experience. All work was carried out by this reports' three authors during January-March 2016.

We thus *do not* offer the solution to the problem of differentiated validation in this report. We also cannot claim confidently that we have covered all aspects of the problem at hand. The problem is too grand and complex to offer a solution or complete overview of all its aspects

with the means at our disposal. Instead we offer a starting point towards a solution; we offer the building blocks for other researchers, policy-makers and perhaps designers/developers to find their own solutions.

We have explicitly limited ourselves to the issue of *validation,* i.e., the study of whether or not claims as to the designed game and app can be empirically substantiated. We have thus not looked further into risk assessment. This is for both substantive and practical reasons; as our desk research progressed we came to the conclusion that concerning validation alone, there is much to be reviewed and discussed. A differentiated approach to validation is already a complex issue. We expect a differentiated approach to risk assessment to be equally if not more complex.

The report is structured as follows. Chapter 2 reviews recent scientific literature of validation of health-oriented serious games and apps. It reports several insightful, instructive example studies and meta-reviews concerning different validation strategies for health-oriented games and apps. Chapter 3 offers a conceptual framework for a differentiated validation strategy. Next, Chapter 4 offers insights into existing tools and initiatives for specific validation strategies (often also addressing further risk assessment). The chapter also presents what new tools or initiatives could be developed as a next step. In Chapter 5 we present the conclusions with respect to the validation of health-oriented games and apps in a differentiated way, as well as what issues for future research these conclusions raise. The Appendix provides the results of a small survey of industry and academic professionals with experience with validation of health-oriented serious game and app.

2 Reviewing relevant validation literature

2.1 Game/app effectiveness

In aid of further study into validation strategies and subsequent validity levels of health-oriented games/apps, we first briefly review the outcomes of the many related studies that have been carried out over the past decade. The focus will specifically be on meta-reviews about the effectiveness of games in general and of games in the health domain. We subsequently highlight two recent case studies where health-oriented games and apps have focused on validation and examined effectiveness.

2.1.1 On the effectiveness of serious/applied games in general

An increasing number of recently published studies, including both quantitative and qualitative meta-reviews, have shown the potential effectiveness of serious games: Vogel, Vogel et al. (2006), Ke (2009), Sitzmann (2011), Wouters and Van Oostendorp (2013), Connolly, et al. (2012), Wouters, et al. (2013) and Clark, Tanner-Smith, and Killingsworth (2015). The overall conclusion of these meta-reviews is that serious games were more effective than conventional instruction methods.

However Connolly, et al. (2012) conclude after analyzing 129 studies that while positive empirical evidence concerning the *effectiveness of games-based learning* was found, RCTs clearly provide more rigorous evidence concerning the impact of games compared to more commonly used quasi-experimental designs and surveys. They recommend that more RCTs should be carried out. A recent update of their meta-review (Boyle, et al., 2016) analyzed 72 papers on serious games between 2009 and 2014 and confirms this recommendation. Boyle et al. (2016) also compared the studies' methodologies (RCT, quasi-experimental, and other) and found only 15% RCT designs, 46 % quasi-experimental, and other designs in 49% of 72 studies, which was similar to the period before 2009. Furthermore, they note that most serious games are focused on knowledge acquisition. Knowledge acquisition varied from topics such as STEM (science, technology, engineering, and mathematics) domains and health.

Indeed most of the previously mentioned meta-reviews are focused on learning outcomes regarding acquired knowledge and skills. One exception is the review of Wouters et al. (2013) who in addition to learning outcomes, also analyzed *motivational effects*. They found that serious games were more effective than instruction methods in terms of learning and retention. However they were not rated as being more motivating compared to conventional instruction methods.

Related to the topic of motivation, the meta-review of Boyle et al. (2012) contained an interesting conceptual analysis on *engagement* and discusses the role of engagement in 55 digital entertainment games between 2002-2012. Their analysis focused on diverse aspects of engagement in games, including subjective experiences while playing games (flow), the physiological concomitant of these experiences, and the motives and motivations for playing games.

It is important to be aware of the fact that a knowledge gap remains in terms of gameplay and intended outcomes. There are many examples of games that have a significant effect on factors such as motivation, learning, and engagement. However it is still unclear how this

outcome is attained. In most cases, for instance, it is unclear which game features are responsible for a specific outcome (Wouters & Van Oostendorp, 2013). This question of the 'construct validity' of a serious game designed as an intervention, is not merely an academic issue. It is also an important practical issue, as determining the decisive features can contribute to the efficient development of a new related game.

2.1.2 On the effectiveness of health-oriented games/apps

We are particularly interested in the design and effects of health-oriented games and apps. In this section we will briefly look at 11 meta-reviews on games/apps in the health domain.

In a now somewhat older, narrative systematic review, Baranowski et al. (2008) use a qualitative analysis comprised of 27 articles to describe the results of 25 video games aimed at promoting health-related behavior change. Behavior change focused on a broad range of topics including fitness, dietary change, coping with asthma, and adherence to medication. They claim that most articles demonstrated positive health-related behavior change due to playing video games.

Two important methods in the game influenced behavior. The first method included inserting behavior change procedures such as goal-setting or self-regulation into gameplay. The second method involved the use of narrative and integrating concepts of behavior change into that narrative. The story included in the game should therefore address the relevant behavioral change in order to influence behavior.

Adams (2010) examined a small number of serious games for health. These games focused on health promotion (e.g. learning how viral infections are spread), prevention, or treatment (e.g. aimed at behavior modification to increase compliance for patients with asthma or diabetes). A number of the reviewed studies indicated positive results. However, Adams indicated that there were also many unknowns about the efficacy of many of the games. The complexity of a health-oriented game design, development and subsequent efficacy study project is the primary reason for the limited number of efficacy studies at the time, according to Adams.

A meta-review study by Rahmani and Boren (2012) examined 54 articles in the period 2000-2012. Only RCT studies were included. Several types of games were analyzed: games on pain and stress reduction, patient behavioral change games (e.g. exercise games, also known as exergames), patient rehabilitation games, diagnostics tools and cognitive ability games. They concluded that, while exergames were most prominent choice regarding health improvement at the time, most of the studies have shown promising results.

Similarly, after qualitatively reviewing many commercially available and tailor-made health games, Kato (2010) concluded that in general these games had a positive impact in areas such as nausea in pediatric cancer, anxiety management, physical therapy, burn pain, diabetes, asthma, bladder and bowel function.

Kueider et al. (2012) examined studies aimed at training domains such as memory, executive functioning, attention, and other cognitive areas via three types of intervention: classic cognitive training, neuropsychological software, and video game interventions. Results indicated that computerized training and game interventions were just as effective and less labor intensive alternatives to traditional approaches.

Cannon-Bowers, Bowers and Procci (2011) discuss approximately 25 video games and corresponding studies where video games were deemed an effective tool for: training healthcare professionals, therapy and disease management (diabetes, respiratory disorders, cancer, rehabilitation), disease prevention, and wellness and lifestyle (which included exergaming). They state that studies should be designed to reveal how and why a game may or may not be effective in achieving its goals. They also state that strong behavioral theories should be integrated into serious games.

Premack, Carroll, McNamara et al. (2012) include 38 RCT studies that evaluated the potential effect of video games on health outcomes.. Health outcomes were illustrated via the use of physical therapy, psychological therapy, disease self-management, distraction from discomfort, and physical activity. They conclude that video games could potentially improve health outcomes, particularly in the areas of psychological therapy and physical therapy. Furthermore, they posit that the use of RCTs will help build evidence in this area.

Ghanbarzadeh, Ghapanchi, Blumenstein and Talaei-Khoei (2014) discuss more complex games. They report studies on the use of three-dimensional virtual worlds (3DVWs) in healthcare. They cover 62 publications from 1990 to 2013, of which nine articles concerned patient treatment and three concerned lifestyle. They conclude that 3DVWs could be of value and offer insights to the healthcare community. Unfortunately outcomes in terms of improvement are not mentioned

An interesting and important article is the meta-review study by Hamari, Koivisto and Pakkanen (2014). They examined 95 studies involving persuasive technology, i.e. games designed for the purpose of guiding the user towards an attitude or behavior change, and found that more than 90% of these were successful in persuading users. A large part (about 50%) of the applications analyzed focused on health and exercise. Nevertheless, Hamari et al. conclude that many studies did not measure experience/engagement and attitudes with validated scales, and some lacked control groups or relied solely on user evaluations.

An extensive meta-review from DeSmet, Van Ryckeghem, et al. (2014) includes a review of serious games aimed at a healthy lifestyle. They analyzed 54 studies and concluded that these games had small significant positive effects on healthy lifestyle, particularly on knowledge, a relevant determinant of behavior change. DeSmet et al. (2014) add that the effect size of serious games on behavior are in line with findings of other meta-analyses on computer-delivered interventions that are not games (Krebs et al., 2010; Portnoy et al., 2008) and conclude “health professionals and policy makers may therefore consider serious games as an alternative to other computer-delivered interventions”. Further research indicated the use of rigorous designs with high external validity, longer play durations, and dynamically adapting the game to player’s game experience and play proficiency.

Subhi, Bube, Bojsen et al. (2015) recently reviewed 52 studies that included 6,520 mobile phone apps. They focused on the prevalence of expert involvement in app development and whether app contents adhere to current medical evidence. The apps relate to many topics: dermatology, ophthalmology, pain management, asthma self-management, prostate cancer, obesity, etc. They concluded that most medical mobile phone apps lack expert involvement and do not adhere to relevant medical evidence.

The above review of games and apps provide us with a general indication of the use of games in the healthcare sector, particularly for disease prevention and health promotion. Although it is apparent that many serious games have successfully demonstrated their effectiveness, more attention should still be paid to the uptake of serious games, i.e. the social processes that connect targeted participants to games, apps and their devices.

2.1.3 Recent empirical example studies on health-oriented games and apps

What becomes clear from the previous meta-reviews is that many games/apps claim to prevent, treat, cure or care a condition. Such a claim is more far-reaching than providing information, raising awareness or learning/training. This again shows the importance of differentiating different validation strategies, i.e., an elaboration of what validation should or could entail in these different cases.

To further elaborate the need for and specifics of validation in these more far-reaching cases, we offer two recent empirical example studies. We have focused on serious games and left out games dedicated to educational aims. Thus we *excluded* games for training of medical professionals or students. Furthermore, we focused on games/apps for the general public/end users, not in a formal educational setting. The first study discusses a serious game on an iPad platform, the second discusses a mobile phone app.

Example study 1: Padua Rehabilitation Tool

The application developed by Cardullo, Gamberini, and Mapelli (2015) concerns the Padua Rehabilitation Tool (PRT), a neuro-cognitive rehabilitation tool for patients with dementia. The PRT consists of a suite of mini-tests, containing 35 cognitive stimulation exercises of grouped by the various cognitive functions involved: attention, memory, language, logical reasoning, recognition, orientation and motor control. The interface used is simple. The exercises consist of levels of increasing difficulty. Visual and auditory feedback is provided.

One study implemented a pretest/posttest RCT design, with two control conditions, a no-treatment condition and a traditional paper and pencil condition. Participants were randomly assigned to conditions. All patients were assessed before treatment, after treatment, and one month after the end of the treatment with standardized cognitive assessment tests.

The results of the study showed that tablet device technology could provide good results for cognitive rehabilitation. Patients in the iPad condition improved more than the no-treatment control condition, though the difference between both treatment conditions was not significant. Perhaps more importantly, the results obtained revealed a high level of appreciation and efficacy from the patients that used the PRT.

The main reason why this game was mentioned as a case study is that it illustrates the effort made to set up a valid research design using a serious game. They used two control conditions, a random assignment which enabled a valid intervention design, and included valid measurement instruments, paying attention to external validity by repeating measurements after a delay.

Example study 2: Alcohol Dependence Intervention

Gamito, Oliveira, Lopes et al. (2014) present an extensive study on the effects of an intervention using mobile phone technology (run on Android OS, using a Samsung Galaxy smart phone) and serious games concerning patients with alcohol dependence.

The cognitive stimulation program implemented in the game comprised attention, working memory, and logical reasoning. The difficulty of the game increased progressively.

The study used an RCT with a pretest-posttest design. The control condition consisted of only the usual alcohol-abstinence program. Mean age of participants was 45 years, N=54, and participants were recruited from an alcohol-rehabilitation clinic. The intervention in both conditions took 4 weeks, on a 2-3 days/week basis. The pre- and posttest consisted of a standardized cognitive abilities test and some other measurements (assessment of frontal brain functioning etc.).

Positive effects were found for the mobile game intervention compared to the control condition, particularly on tests reflecting cognitive functioning of the frontal lobe. Our impression is that important effects can be found also in the realm of eHealth apps. The study is methodologically thorough.

2.2 Concerns

We have introduced many meta-review studies based on a substantial amount of individual studies. The general picture of the effects of serious games, also in the area of health, is positive. It is important to stress that we did not leave out *negative* meta-reviews. Still, the above overview generates a few concerns. Below is a summary of our main concerns.

Validity:

- Validated measuring instruments (measurement tools) are needed, e.g. on the area of measuring experience/engagement and attitudes;
- A recurrent comment is that more RCTs should be carried out, an issue of internal validity, i.e., an issue of how well the experimental research design allows the researcher to test the hypotheses at hand;
- More research is needed on the specific game features to determine their effectiveness, an issue of construct validity;
- Attention should be paid to external validity (other settings, other groups of participants, other time frames, e.g. longer play duration);
- It is wise to involve medical experts in developing serious health games and apps.

Game Design:

- Elaboration of the motivational processes and effects of game play is needed;
- Somewhat more detailed, the game should include behavior-change procedures, preferably in the context of a game story;
- Dynamic adaptation to the game experience and proficiency of the player is needed;
- Attention should be paid to the social processes that connect participants to the devices, so that participants integrate them into their daily life.

3 A conceptual framework for health-oriented game/app validation

With a firm grasp of relevant scientific literature on health-oriented games/apps validity and validation, we can dive into the problem of differentiated validation much deeper. We do this by first exploring the relevant concepts in this chapter.

3.1 A goal/intervention categorization

As Renger, Veltkamp and Schouten (2015) also noted, an exploration of differentiated validation should start with a better understanding of the different types or categories of goals and interventions that health-oriented games/apps pursue. For instance, an app intended to disseminate information about cancer screenings would be subject to less rigorous validation procedures than an app designed to detect potential cancer symptoms and direct the end-user to a healthcare professional if needed. The categorization presented below is based on game/app types covered by the literature reviewed in Chapter 2.

Concerning medical professionals (doctors, surgeons, etc.) or even the general public, we see games/apps aimed at:

a. *Learning* (explicitly/consciously obtaining information or gaining knowledge e.g. concerning resuscitation or the human body);
b. *Training* (explicitly/consciously developing and practicing competencies – the combined use of knowledge and skills, e.g. first aid, trauma treatment or laparoscopic surgery).

Concerning patients the following typology is relevant:

a. *Prevention*: hindering a condition from occurring - fitness, dietary change, healthy lifestyle and common disease prevention measures. This could (also) involve learning, but not necessarily.
b. *Diagnosis*: measurement of patient status variables (directly or through self-assessment), followed by analysis and determination of one or more (potential) conditions – depression, respiratory disorders, diabetes, (a form of) cancer.
c. *Treatment or cure*: relieving a condition partly or completely – asthma or other respiratory disorders, diabetes, bladder and bowel function, physical therapy, rehabilitation, medication adherence, treatment compliance. This could (also) involve learning or training, but not necessarily.
d. *Care*: alleviating (symptoms of) a condition, or halting the progression of a condition, including those that come naturally with old age – memory training, attention training, executive functioning, pain and stress reduction, nausea relief in pediatric cancer, burn pain, anxiety management. This could again (also) involve learning or training, but not necessarily.

It is important to stress that in the case of prevention, treatment/cure or care oftentimes the objective/intervention is behavioral in nature. How games/apps (try to) achieve the desired behavior or a change in behavior differs. They might first try to raise awareness of an issue by offering directly or indirectly information/knowledge (e.g. in a suggestive story offered throughout the entire game/app experience). They might require and assess actual behavior in

the design itself, oftentimes procedurally, offering positive or negative reinforcement along the way.

3.2 Relevant outcome variables

For an intervention to meet all its intended goals, multiple elements need to be in order.

We posit that the main dimensions necessary to validate any game or app are: usability, playability, efficacy and side-effects. Each of these dimensions apply to a greater or lesser extent to a specific program (game or app), and so the instruments need to be tailored to them in order to validate those specific programs.

Usability. Usability is concerned with the functionality of the program. Are users able to do what they want with the program? Are they able to use it as the creators intended? While usability can be seen as a very broad topic of human-computer interaction, we define it here as functionality and accessibility. An app with powerful persuasive effects cannot exert such effects if users cannot figure out how to get it to work. Usability is the combined whole of the intended usage, presentation style and technological back-end.

Usability cannot reliably be self-assessed by the developers. The program ultimately needs to be tested with the target audience, preferably in a natural setting. Usability can be affected by many different criteria such as things as battery life and readability of the tutorial.

Playability. Playability is related more to the experience of interacting with a program than it is about the functionality of the program. Does it give users freedom to express themselves or act as they please within certain boundaries without serious consequences? While usability is about making sure the program works and is suited to its targeted audience, playability ensures these users are having fun with the program. Although a lack of playability will not necessarily stop users from using the program, it will prevent them from getting the most out of the experience and coming back to use it again.

Playability is an elusive concept that is usually measured through its effects on user's mental state, i.e. it is, like usability, not something developers can determine without letting others experience their program. Measurement can consist of immersion or presence scales (Olson, Procci & Bowers, 2011), as well as indicators of enjoyment.

Efficacy. Efficacy describes the ability of a game or app to shape attitudes, knowledge, skills and/or behaviors of its users towards an intended state. For instance, if a game intends to make players think about their caloric intake, it can be said to be successful if a player starts checking the nutrition labels on food items (not having done so previously). Here, the developer's intentions are integral to deciding what the efficacy of a program could be.

Efficacy can be as simple as making someone aware of something or as difficult as permanently changing someone's behavior. Therefore we cannot make any statement about how this can be measured *in general*, other than the assertion that measurement through end-user self-assessment should be avoided if possible. Behavioral measurements are key here.

Side-effects. Side-effects, on the other hand, encompass all explicitly or implicitly unintended effects a program has on its users. If the previously described game about monitoring caloric intake would cause some of its users to develop eating disorders, this would be an unintentional, adverse effect. The ideal game therefore has good efficacy while limiting side-

effects. Due to the wide scope of side-effects, measuring these dimensions of a program, involve selecting measurements on a program-by-program basis.

Efficacy measurements should be selected based on the intentions of developers. In contrast, side-effect measurements could be selected or developed through previous research, expert evaluations of the program, or from public fears about negative effects.

3.3 Relevant validity types

In Chapter 2 we already mentioned different forms of validity. We elaborate these concepts because they are integral to our suggestions about how to ascertain validity of health-oriented games/apps.

It is useful to make a distinction between games/apps as:

1. *Measurement instruments* that aim to provide information about assessment (self-diagnosis, monitoring, etc.) of acquired knowledge, a skill or attitude or of a certain condition;
2. *Treatment* of, or *therapy or intervention* for a certain condition.

For the first category psychometric characteristics are important. In general, it suffices then to examine the (cor)relation with other measuring instruments.

For the second category more specific information is needed, i.e. information about the causal character of the treatment in relation to its effect is needed. This is the case when the researcher or developer has in mind that the treatment is better than something else (another treatment or just maturation).

In any case, when it comes to validity we recognize and follow the seminal works of Campbell and Stanley (1966) and Cook and Campbell (1969), as well as subsequent sizable contributions by the American Psychological Association, and simulation-/game-specific contributions by several authors that we review in the following section.

3.3.1 Validity concerning games/apps as *instruments*

When it comes to games/apps as instruments for assessment, we recognize the relevance of these instrument validity types:

a. *Content validity* focuses on the validity of the knowledge, skills and/or attitude that the game/apps should aid in obtaining. Is the content complete, correct, and nothing but the intended focus of the game/app?
b. *Face validity* focuses on the validity of the strategy that the game/app proposes for attaining its goal at face value. Do relevant experts (game/interaction designers, behavioral psychologists, educators) recognize the proposed features, mechanics or dynamics as suitable to reach the intended goal?
c. *Construct validity* focuses on the validity of the chosen way of assessing the goal itself. Is the intended goal adequately measured after playing the game/app?
d. *Concurrent validity* focuses on the validity of the game's/app's design in reaching its goal compared to other methods that have also proven to be able to reach said goal. Can the game prove itself as a viable alternative?
e. *Predictive validity* focuses on the validity of the game's/app's design in reaching its goal in a multiple, different settings and situations, including outside the game/app

itself. This comes very close to *external validity,* i.e., the extent to which the results of the validation study can be generalized to other situations and to other people.

The relevance of the above validity types was recognized by Feinstein & Cannon, when modern-day serious game design was still in its infancy, while simulation game design had already peaked.

Mayer et al. (2012) offer a methodology for research into serious gaming for advanced learning, i.e., learning to identify and deal with the complexity of professional practice. The methodology is based on the need to work towards a generic evaluation/assessment framework to allow for comparative, longitudinal research. Such a framework allows serious gaming to develop as a scientific field. The authors offer and call for quasi-experimental research designs involving pre-test, in-game/during-game, and post-test measurements.

They also call for the use of standardized measurements with *construct validity*, obtained using a 'scoop model' of appropriate data gathering methods. The actual variables that need to be gathered concern the pre-game condition, quality of the game design, play and facilitation, the post-game condition, first-order learning, second-order learning, and finally sets of background, mediating and context variables. By doing so Mayer et al. help design researchers of serious games for learning purposes set up validation studies with adequate internal validity.

Further work into the relevance of the above validity types for serious game design in general, and within the domain of health in particular, comes from Graafland et al. (2012, 2014) as well as Warmelink et al. (2015). Following an extensive review of serious games for health (2012), Graafland et al. came to similar conclusions as Mayer et al. (2012), pleading for a more systematic evaluation/assessment methodology. Contrary to Mayer et al, however, Graafland et al. based their methodology on both empirical and non-empirical assessment, as well as the differing roles and natures of validity in the empirical assessment. They also broadened the scope by focusing on for instance data protection for user privacy.

Warmelink et al. (2015) used Graafland et al.'s (2014) validity types to explore practical strategies and associated tools that serious game designers could use *during* the entire design process, rather than solely at the end in the form of an evaluation or assessment. Based on an iterative design process model, they argue that *content validity* can already be pursued in the very first stage of design, when design requirements and the game's objective and content are defined. *Face validity* can be pursued when subsequent concepts are developed by e.g. involving external game design reviewers and subject matter experts, and/or juxtaposing the concept onto validated game design principles or patterns. *Construct validity* can be pursued by doing multiple play-tests of game prototypes in a quasi-experimental fashion. *Concurrent* and *predictive validity* can be pursued when late prototypes or the (nearly-final) game product is integrated in its intended setting through more elaborate (quasi-)experiments, RCTs or continuous 'stealth assessment' through analyses of automatically logged gameplay data. By incorporating these five validity types *into* the design process, designers can spread their resources, make deliberate decisions about what validity strategy to pursue, and limit their dependence on single evaluation/assessment at the end of a costly design/development process.

3.3.2 Validity concerning games/apps as *treatments (interventions)*

In general, in drawing conclusions on effects of specific treatments or interventions, such as playing a serious game, we have to distinguish different types of intervention validity: the statistical conclusion validity, internal validity, construct validity and external validity.

a. *Statistical conclusion validity* concerns the relationship between the treatment and the outcome. Is there a statistically significant relationship? E.g. is the experimental (game) group really better than the control group, or is there really a progress in the experimental group after playing the game?
b. *Internal validity* concerns the causality of the relationship, and that the relationship is not a result of a variable that has not been measured or that we have no control over. In other words, does the treatment cause the effect?
c. *Construct validity* is concerned with the relationship between theory and observation. If the relationship between cause and effect is causal, we must ensure two things: 1) that the treatment reflects the construct of the cause correctly and 2) that the outcome reflects the construct of the effect correctly. So in the domain of games: what are the effective components (or features) of the game that are responsible for the desired effect?
d. *External validity* is concerned with generalization. Can the results be generalized outside the scope of the study? For games, for example, can we expect to find the same results with other participants, other settings or other application domains?

This framework, including the relevant threats of different types of validity (such as low statistical power, reliability, history, selection, mono-operation bias, interaction of selection and treatment, etc., etc.), is more recently discussed by Gray and Salzman (1998) and nicely summarized in a volume by Wohlin et al. (2000), which we used here.

In general, 'true' experimental designs like RCTs – consisting of the random assignment of participants to conditions, objective manipulation of the independent variable(s) while holding (statistically) constant other influences - are necessary to uniquely attribute the presumed cause to the measured effect of games and to exclude alternative interpretations.

Many authors point to this necessary condition for internal validity while discussing empirical game studies. We will mention some conclusions with respect to this issue by some well-known authors in the serious games domain.

In their review of video games as healthcare tools, Cannon-Bowers, Bowers and Procci (2011) recommend studies that reveal how and why a game may or may not be effective in achieving its goals. Investigating simply whether a game is effective or not, does not provide useful insights about how to design such a game, or to generalize results beyond the current application. Furthermore, strong (theoretically based and empirically verified) behavioral theories must be used as a basis to design the serious games. Only in this way can a true science of games in healthcare emerge as results are accumulated.

In addition, Kato (2012) claims that our standards for validation of the growing number of games need to be raised. If not, the games will be overlooked and disregarded. Studies need to be conducted that allow to verify a causal link between playing the game and the outcomes (Kato, 2010).

Finally, Van der Kooij et al. (2015) discuss the way in which the control condition in for instance an RCT, should be formed. They distinguish three options: a) a waiting list control condition, b) a treatment as usual condition, and c) a placebo control condition. They favor the last option because this condition is equal to the investigated serious game in all elements *except* the presumed feature responsible for the effect. They call this game feature the "change catalyst".

An example here is the study by Soekarjo and Van Oostendorp (2015) on evaluating the effect of a persuasive serious game focused on players' energy reduction. They used a control group that received the information as similar as possible compared to the play condition, in a PowerPoint presentation that contained slides with pictures, layout etc., similar to the game. In this way they controlled – as much as possible – for medium, look-and-feel and social aspects and relevant factual information.

Apart from concerns about the internal validity of games as treatment or intervention, it is equally important to examine exactly what component(s) of the treatment were responsible for the effect that was found. This involves the issue of *construct validity*. In a recent review Ke (2015) states that next to the evaluation of the empirical results of games, focus of research should (also) be on providing a detailed record of game design features and process. Evaluation should contain elaborated theoretical underpinnings, design strategies and design rationales for game mechanics and the game world design.

In other words, future game research should pay more attention to the analysis of game features and the way these features become effective. This aligns with the opinion of Mayer (2011) that a 'value-added' approach to game research is needed. It also confirms the conclusion of Boyle et al. (2016) that we now need systematic exploration focusing on which game features are most effective in promoting engagement and supporting learning.

All these suggestions highlight the sense of urgency pertaining to acquiring more information about the construct validity of serious games, in addition to the statistical significance and size of effects (i.e. statistical conclusion validity) and the exclusion of alternative explanations (i.e. internal validity).

3.4 RCTs and other study designs

3.4.1 RCTs

When we want to ascertain the efficacy of an actual *intervention*, we essentially want to test a hypothesis, i.e., a causal inference between two variables. In pursuit of such inferences, research often defaults into the RCT design (Ioannidis, 1998). RCTs are seen as the gold standard of research design practices, especially in medical science and social science, though the terms applied may differ. However, it is by far not the only means available. For e-health applications and games, we can and should consider far more than the RCT research design. Different experimental research designs should be considered as next-best options.

Within the field of experimental research design the RCT is actually only a framework for research design rather than a full-on research design itself. An experiment can be considered a RCT when it fulfills three conditions (Bonell et al., 2011, p. 582):

1. *There is or will be a control group*, i.e., a group of participants who will not get or undergo the intervention, but instead a currently accepted alternative to the

intervention (experimental group). *The comparison of the experimental and control group aligns with the independent variable*, i.e. the objective manipulation of the variable of interest by the researcher while holding (statistically) constant other influences.

2. *Assignment of participants* to the intervention or the control group (individually or clustered) is *completely randomized*. To ensure randomization, one should use full-on random selection software or protocols. Treatment allocation might be (but does not have to be) hidden from participants, providers and/or researchers, to avoid potential impact of having this information on the outcome variable.
3. The *outcome variable used in the hypothesis is measured in both groups* in the same manner and at the same time.

An RCT is often, but not necessarily, a form of quantitative research, thus quantification of variables and a sufficient number of participants to subsequently be able to carry out valid statistical analyses are the basics of RCT design as well.

Still, the above is only a framework for design. There are many forms of RCT designs. Each RCT will require the researcher to subsequently find out: 1) how and when to validly measure the outcome variable; 2) what confounding variables should and can be validly measured; 3) how many participants should and can be recruited and subsequently; 4) how the involved researchers and participants should be instructed and supervised; 5) how participant attrition/dropout can be prevented. When answering each of these questions, different forms of biases and techniques to limit those biases must be considered.

Although it is beyond the scope of this report to go into specific design issues, more in-depth, we will end this subsection with the general distinction in health efficacy study designs presented by Victora et al. (2004), based on different kinds of research objectives they identified:

- '*Clinical efficacy trials* … follow the standard clinical trial model in which study participants are individually selected and randomized and the dose is ensured at the individual level. … To ensure ideal compliance, staff in clinical efficacy studies are intensively trained, supervision is strong, subjects are intensively counseled and may be reimbursed for any expenses associated with the intervention (e.g., transportation to clinic), dosages are strictly controlled, side effects are monitored and managed, and non-compliers are actively sought' (Victora et al., 2004, p. 402). These are truly laboratory experiments, and so are not naturalistic.
- *'Public health regimen efficacy studies* … are similar to clinical efficacy trials, but the intervention is applied to groups rather than individuals. The optimal dose of the intervention is delivered to every subject and compliance is ensured' (Victora et al., 2004, p. 402).
- *'Public health delivery efficacy studies* … like regimen efficacy studies, ensure that an optimal dose of the intervention is delivered to the individual or family. However, there is no resource-intensive effort to promote compliance, although compliance is likely to be somewhat above that observed in routine circumstances (and is thus described as "best practice")' (Victora et al., 2004, p. 402).
- *'Public health program efficacy studies* … entail making the intervention available to the health services but not promoting any resource-intensive efforts to ensure optimal

delivery or compliance. Thus, behavioral factors pertaining to health systems and individuals are allowed to affect the dose of the intervention. Given the presence of the study team, delivery and compliance are likely to be somewhat above routine levels, described here as "best practice"' (Victora et al., 2004, p. 402-3).

- *'Public health program effectiveness studies* … entail allocating geographic units to receive or not receive the intervention but making no additional efforts to improve delivery or compliance above routine levels' (Victora et al., 2004, p. 403).

3.4.2 Alternatives to RCTs

There are several reasons for not being able to use a RCT research design. Some of these are related to the nature of the intervention under study and the three main characteristics of RCTs. For example, in many cases pure randomization of participant assignment can be practically impossible to implement or unethical given the outcome that the intervention is meant to achieve (think of a cure for cancer for example). It could also be argued that RCTs tend to take a lot of time. Indeed, Ioannidis (1998) found that RCTs of evaluating the effects of medical treatments took a median of 5.5 years, from design to publication.

Research into next-best alternatives to the RCT design have yielded several designs that can be employed when one or more of the three conditions mentioned above are not met. We will discuss a few examples here (in no particular order):

- *Stepped-Wedge Trial*, when conditions 1 and 2 cannot be fully met. The intervention is rolled out to all participants (either individually or clustered) over a number of time periods, step by step. The order of interventions received is randomly allocated. Data is collected every time a (number of) new participant(s) receives the intervention. As Brown and Lilford put it, a stepped-wedge trial is particularly useful 'if there is a prior belief that the intervention will do more good than harm' (Brown & Lilford, 2006).
- *Crossover Trial*, again when condition 1 and 2 cannot be fully met. All participants receive the intervention for a predetermined period of time and the control condition for a predetermined period of time. The order (intervention-control or control-intervention) is determined randomly. As Bonell et al. put it, a crossover trial is particularly useful when 'evaluating acute effects and in scenarios where it is acceptable to withdraw interventions after a period of delivery' (Bonell et al., 2011, p. 584).
- *Randomized Encouragement Designs*, when condition 2 cannot be fully met. A random selection of participants are offered the opportunity to receive the intervention or to go for the control condition. The decision is up to the participants, although they are strongly encouraged to go for the intervention. As West et al. put it, the randomized encouragement design is particularly useful 'for interventions for which it is impractical or unethical to require adherence or in which the necessary incentives would be unrealistic, thus precluding generalization to practice' (West et al., 2008).
- *Nonrandom Quantitative Assignment of Treatment*, when condition 2 cannot be fully met. In this case participants are assigned to the intervention based on a quantitative measure, 'often a measure of need, merit, or risk' (West et al., 2008). A certain threshold on the assignment variable thus forms the tipping point between control condition and intervention. Thus, assuming the experiment was designed and executed well, a difference in outcome variable at the assignment variable's threshold would

roughly indicate an intervention effect. As West et al. put it, the nonrandom quantitative assignment of treatment is particularly useful if 'randomization is impractical or unethical' (West et al., 2008), especially when the reason why randomization is unethical follows a quantifiable predictor variable (e.g. risk of the disease in question) that can function as the assignment variable.

- *Observational Study*, when conditions 1 and 2 cannot be fully met. Groups of participants receive various interventions, including the one under actual study, often by their own choosing. Researchers might base the intervention chosen for a participant (group) on certain confounding variable scores, while participants might do the same, only inadvertently. Thus numerous relevant confounding variables need to be measured and taken into account to aim for valid analyses. Researchers have higher chance at more valid analyses if they calculate *propensity scores*, representing the predicted probability that a participant will receive the treatment given his or her baseline measurements. In the words of West et al., 'If the researcher can accurately construct propensity scores that balance the treatment and control participants on all potentially relevant baseline variables, the difference between the response in the treatment condition and the control condition (conditioned on the propensity scores) will be an estimate of the causal effect.' (West et al., 2008). Researchers can also incorporate certain elements into their research design to aim for more valid analyses, e.g., multiple pre- and/or post-intervention measurements and using multiple nonequivalent outcome variable measurements.
- *(Quasi-Experimental) Interrupted Time Series Analysis* or *Process Evaluation*, when conditions 1 and 2 cannot be met. One could argue that when a researcher does an observational study with multiple interventions including the one under study, and multiple pre-/post-intervention measurements with an adequate framework of confounding variable measurements, he/she is actually doing a quasi-experiment or process evaluation. When the researchers do multiple pre-intervention *and* post-intervention measurements at set times for all participants, he/she is doing an interrupted time series analysis. The arguments for doing such studies/analyses are very similar, if not exactly the same, as those mentioned above under observational study (Bonell et al., 2011).
- *General Population Comparisons*, when conditions 1 and 2 cannot be met. When only one group of participants can be formed, and they all undergo the intervention, a researcher might still be able to obtain data comparable to data already available about the general population from other sources such as (inter)national statistics agencies. That way the intervention group data can still be juxtaposed with the general population data to estimate an effect. Of course, this strategy can only be usefully applied if a sufficient range of valid outcome and confounding variable scores of the general population under question can actually be obtained (Bonell et al., 2011). It will be clear that the internal validity of a study following this design will not be high.
- *Continuous Evaluation of Evolving Interventions*. Over the course of a workshop on evidence of *mHealth* assessment and intervention efficacy, Kumar et al. (2013) shared several designs that fold together previously separated elements. The *continuous evaluation of evolving interventions* method (Mohr, Cheung, Schueller, Hendricks Brown, & Duan, 2013) describes a process of including iterations of interventions whenever they become available. This method allows researchers to find

the most effective strategy through continued comparison by breaking the rule that prevents alterations to research designs after data-gathering has started. Other research designs break down interventions into their constituent parts, separating them by preliminary findings before performing confirmatory studies on more readily observable chunks (Collins, Murphy, & Strecher, 2007). Rather than testing these components in isolation, they are staggered across several conditions, with each condition offering a unique combination of factors. An anti-smoking app, for instance, can modulate the number of reminders it gives its user per week while simultaneously including different encouragement methods. This method is especially suited to interventions with properties that can be tweaked or used modularly. It can also help to generalize results from one intervention to others that employ the same component (e.g. daily reminders).

All of the above research designs can be greatly strengthened by utilizing the relationship with data that is inherent in the apps and games discussed here (Kumar et al., 2013). The continued trend towards always-on devices, wearables, and persistent data logging enables researchers to delve into a vast amount of data generated for each user. By applying these data sources in novel ways, measurements that were impractical or impossible to collect are now freely available. Such processes require automated data-processing and analysis tools to arrive at nuanced and personal histories of use that surpass laboratory measurements in their richness. However, this greater repertoire of measurement data does not automatically solve the validity issue.

What most of the research designs have in common is that they tend to remove obstacles in applying an RCT design. Still, many such designs take prodigious time investments in the traditional cycle of research, sometimes exceeding that of regular RCT designs such as, for instance, the quasi-experimental interrupted time series design.

We posit that the time required is not necessarily a very good reason for not choosing an RCT or ‘spinoff’ design to evaluate a game/app. If designed and implemented by experienced professionals, an RCT in order to evaluate a game/app can be completed within periods that are much smaller, depending on the particular study objectives. The time argument in itself is not a very meaningful consideration for choosing an RCT design or not, irrespective of the loss of validity.

Another concern is that the needed statistical analysis techniques for these alternative designs are often more complex and time consuming than techniques needed for regular RCT designs. Also the required advanced expertise is often not available outside universities.

4 Tools for validation

Having developed a conceptual framework of health-oriented game/app assessment and validation, in this chapter we offer a review of existing validation tools. This subsequently leads us to considerations for new tools for differentiated validation.

4.1 Heuristic evaluation instruments for usability and playability

Olsen, Procci and Bowers (2011) advocate to assess serious games using three aspects: *usability* (independent functionalities within individual components of a system; ease of use of the game), *playability* (concerns enjoyable interaction with a game) and *learning* (efficacy in attaining learning outcomes). They did not develop new instruments for measuring these three aspects. Instead, they referred to existing instruments, mostly known scales like the System Usability Scale (SUS) or the Questionnaire for User Interface Satisfaction (QUIS). They suggest to apply these instruments during the development process of the game, even in the beginning of the process, e.g. during story-boarding and paper prototyping, eventually completed by thinking-aloud protocols when the game is getting more developed.

Desurvire, Caplan, and Toth (2004) developed a comprehensive set of heuristics for game usability, game play, game story and game mechanics. The last three subsets or dimensions can be captured by the general term game play. In total 43 heuristics were distinguished (31 heuristics reflecting playability – e.g. related to game story "the Player has a sense of control over their character and is able to use tactics and strategies", and 12 reflecting usability, e.g. "The interface should be as non-intrusive to the Player as possible" based on notions as feedback, consistency, etc.). So this approach offers aspects in which games can be rated, mainly focused on game usability and game play. Evaluators applied the heuristics on a game in development, and compared the results with more conventional user testing with prospective users. The results indicated there was much overlap between the issues found by applying the heuristics and the user study. Their analysis was not specifically on the domain of serious games, which explains why the learning aspect was not included.

Pinelle, Wong and Stach (2008) analyzed the reviews of many (108) games, and extracted from them in total 285 problems, and categorized these into 12 common classes of usability problems. They developed 10 usability heuristics based on these problem categories. In a pilot study evaluators had to inspect a game and to identify instances that did not match with the 10 heuristics. They concluded that the heuristics are particularly helpful in early game design. They also pointed out that the results did help to identify game-specific usability problems that can easily be overlooked otherwise.

Moreno-Ger et al. (2012) tried in their approach to capture different aspects of interaction of participants with a serious game: learning, engagement and the appropriateness of the design. The last aspect perhaps most resembles conventional usability issues. Their approach, Serious Game Usability Evaluator (SeGUE), evaluates a serious game along two orthogonal dimensions:

a. *System-related dimension*, consisting of 6 categories: Functionality, Layout/UI, Game flow, Content, Technical Error, Non-applicable
b. *User-related dimension*, containing 10 event categories: Learning, Reflecting, Satisfied/excited, Pleasantly frustrated, Frustrated, Confused, Annoyed, Unable to continue, Non-applicable, Suggestion/Comment.

The approach is based on evaluators who observed and analyzed play sessions of real test users and annotated significant events in terms of above 6x10 categories. So far, clear results of the usefulness of their approach are unavailable though they report that a case study helped them to improve a serious game and also served to improve the SeGUE instrument.

Summarizing, we have described several approaches to obtaining insight in the dimensions of usability and playability of games. These approaches were based on expert inspection by (expert) evaluators who identified the problems (based on heuristics) and users experience collected whilst playing the game. The learning aspect was not included here, but many measures (recognition, rating scales etc.) can be borrowed from the area of serious games focused on learning. Further research has to indicate what the most valid and practical method is to obtain reliable usability and playability information for a serious game.

4.2 (Self)Assessment initiatives

Because of the surge in popularity of health-related games and apps, only a fraction of the products currently released are subjected to a rigorous study of their effectiveness. To stand out in this increasingly crowded field and afford legitimacy to, for instance, the medical self-help interventions they developed, content creators are looking for ways to assess apps that minimize the time necessary to obtain a result.

However, the current knowledge base does not allow developers to demonstrate effectiveness of an app purely based on the construction of a game or app. For this reason, Albrecht (2013) calls for a mechanism that allows developers to partially validate their apps by providing transparency on what the intervention is intended to do, what its mechanisms are based on, and how it treats user data. We will discuss several examples of what this would amount to in practice.

4.2.1 Government initiatives

Commissioned by the Dutch Ministry of Health, Welfare and Sport, the Trimbos Institute (focusing on mental health and addiction) has developed an assessment tool-set that aims to provide a sense of reliability to online self-help programs. This service, called the *Onlinehulp Stempel* (https://www.onlinehulpstempel.nl/home) places program developers into a track consisting of three steps.

First, developers fill in a self-assessment focusing on *efficacy, transparency*, *user-friendliness*, *accessibility* and *data security*. The self-assessment separates questions into required elements and added benefits. *Efficacy's* required elements are that the intervention is based on treatments that were proven to be effective, and that members of the target audience were involved in the design process. Although empirical results proving the intervention's efficacy are seen as a bonus, it is not part of the required process. *Transparency* demands include informing possible users what the product is like and who are behind it. *User-friendliness* must-haves are all related to moderating communication in the program and allowing users to contact the developers, while *usability and user acceptance tests* are seen as added benefits. *Accessibility* requirements naturally relate to how broadly the program can be implemented, while the *data security* demands that providers use encrypted servers that are located in Europe to store data. Primarily, this self-assessment is meant to show the intentions of the program developers, asking for a systematic approach to intervention design that is based on current knowledge.

Second, after the intervention has been demonstrated to match the basic requirements, developers can request *expert reviews*. These experts are active healthcare professionals that have implemented similar interventions in the past. The review does not include testing the intervention among target audience members or others. Upon getting a positive review from two experts, the program receives the online help stamp to allow them to advertise their product.

The third step relies on the program's *end users*. They can rate their experience with the program. All such ratings are aggregated as a final end report which is then visible as a grade out of ten on the stamp next to an indication that the program was 'approved by mental healthcare professionals'. Looking into the end user reports, it is clear they are treated similarly to consumer product reviews. Reviewers comment on different elements, and are only asked for a single grade.

The three steps to obtain one of these stamps are prioritized towards three ends: the experience of the end user, the security and protection of the end user and her data, and lastly the intentions of the developer towards (and face validity of) the intervention. Despite being correlated to efficacy (Yasini & Marchand, 2015), the first two of these steps in no way demonstrate efficacy. Instead, the mentality seems to be that as long as the program does not harm user's privacy and offers a good experience, the effectiveness of the program in changing attitudes and behaviors is secondary. This could partly be because of the time and finances required to find empirical proof of a program's effects, but it can also be seen as the product of an industry-wide focus on innovation instead of on prevention and self-curing efforts.

Considering the tremendous influx of new apps, games, and online programs in this field, validating programs could lead to the bubble bursting (rather than sustaining current growth) as only a handful of programs can prove positive effects – and disprove harmful, unintended ones.

Recently the Dutch Municipal Health Services (GGDen) collectively took the initiative to start an app store where mobile e-health applications are collected (https://www.ggdappstore.nl/), centrally judged, and given a vignette ('keurmerk'; consisting of 0-5 stars). The apps, now approx. 50, are described using two dimensions:

1. Description (purpose, target group, for whom?, to be used for?, functionality, where to find?)
2. Rating (usability, reliability, evidence, privacy). These 4 categories are qualitatively judged with ratings as (in)sufficient, adequate, etc. and accompanied by a brief explanation.

The apps are described in a simple and understandable way, but it is not exactly clear what rationale is used for giving the vignette rating.

In Great Britain, a list of health apps was curated online by the National Health Service (Boulos, Brewer, Karimkhani, Buller, & Dellavalle, 2014). The website allowed developers to submit their interventions to be reviewed by curators. The repository linked through to online market places instead of hosting the content, and did not appear to validate the programs. Efforts appear to be underway to increase the scope of this site to become an endorsement

site for particular apps (Boulos et al., 2014; NHS Choices, 2015). The legitimacy of this sort of hosting service could help those who abide by these guidelines to stand out from the crowd.

Alongside this project, the British and United States' governments are getting to grips with the different kinds of apps and games that are being published. The US Food and Drug Administration is classifying apps on their intended use (Boulos et al., 2014), even endorsing certain apps (Cain, 2012). An ongoing discussion is being held on how to separate programs that require this endorsement from those with less invasive interventions. Very likely, this categorization will be along the lines detailed in Chapter 3 of this article by being based on the intended usage.

Similar efforts have, back in Great Britain, led to the use of a seal of approval by the Medicine and Healthcare Products Regulatory Agency (MHRA, 2014). This places the burden of evidence on the developer of the program and requires them to provide evidence for its efficacy and lack of unintended effects. This requirement is the likely reason that up until 2013, no developers could apply the seal to their program yet (Genetic Digital, 2013).

4.2.2 Academic initiatives

Similar processes are taking root in different sectors. In academia, the Journal for Medical Internet Research's mHealth branch has been working on a peer-review tool for mobile apps (JMIR, 2012). Like the *Onlinehulp Stempel*, this tool would bestow a 'transparency seal' to developers keen to advertise their apps that is linked to information about the app (upon developer disclosure) as well as to ratings of peers and/or end users.

While currently still in its infancy, the developers of this tool hope to complement these two assessment criteria with research-based usability and usage results as well as health outcomes through RCTs. How these latter elements will come to be implemented is currently not known.

Graafland, et al. (2014) discuss a framework for the assessment of specific medical serious games. Their framework provides 62 questions in 5 main themes aimed at assessing serious games. The themes are rationale, functionality, validity and data safety, next to a description of the game's theme. The framework should allow caregivers and educators to make balanced choices. Furthermore, it should provide game manufacturers with standards for the development of new, valid serious games.

In our view, how balanced and useful the 62 items are, in terms of their content validity, remains to be seen. For example, the playability aspect seems to be missing. Furthermore, it contains no information (or question) on the efficacy of the game or (internal or construct) validity of the game as an intervention (when applicable).

Following their publication, many, if not all, of the authors instigated the Quality Label for Serious Games in Medicine as a service offered by the Dutch Society for Simulation in Healthcare (DSSG; http://www.dssh.nl/en/quality-label/). This Quality Label can be requested by any and all health game developers. If the label is awarded, the health game in question will receive a number of stars (depending on the committee's judgment of the information provided) and a report. The idea is, of course, that by getting a Quality Label by an independent third party, a health game will be accepted more quickly and more broadly by the

involved stakeholders (e.g. patients, health providers, insurance companies). At the time of writing, 11 health games have been awarded the label.

A health game developer wishing to obtain the quality label first fills in a form that follows many of the items and questions by Graafland et al. (2014). The aforementioned distinction between face, content, construct, concurrent and predictive validity form the foundation of the validity assessment. Details concerning with which study (experimental or otherwise) each kind of validity is exactly ascertained are not explicitly asked for. Many additional items/questions of the form concern aspects of validation, for example:

- 'What will the player learn and up to which level?'
- 'Which parameters are (to designers' opinion) indicative for measuring learning effects?'
- 'Please describe restrictions and limitations of the serious game.'
- 'Please state potential undesirable effects.'
- 'Did user testing take place? What were the learnings, and how were these incorporated in the game?'

Subsequently a committee of medical specialists from different centers in the Netherlands (many of whom have experience with health games) review the supplied data. They draft their report and decide whether or not to award the label, and if so, with how many stars.

4.2.3 Issues and critiques

Though they are met with enthusiasm, the aforementioned assessment projects are not achieving major results so far. The *Onlinehulp Stempel*, for instance, has processed two programs at the time of writing, both of which received around 10 reviews by users, while the JMIR tool has to our knowledge, not yet been launched. We will discuss several possible explanations for this lack of use.

First, as Albrecht (2013) mentioned, assessments that are not directly visible in the market place or app store where the programs are sold, will not be of great use. For example, prospective users might need to sift through a large number of competing alternatives before finding an app, which limits the information-gathering they would do on each to look at the in-store descriptions. It might put developers off from investing time in such a tool if they cannot apply it to where their audience might see it.

Second, another issue likely prevents more program developers from engaging in self-assessments. From Subhi et al.'s (2015) review of expert involvement and adherence to medical evidence under a total of 6502 medical smartphone apps, it is clear that the majority of these apps are not congruent with the medical evidence they are supposed to build on. Moreover, many apps do not provide an indication that a medical professional (or expert) was involved in its creation.

Most of the apps under study therefore do not adhere to any standards that would allow them to apply the *Onlinehulp Stempel*, and one could say that being transparent about their origin and intentions is not in their best interest. Seen in this way, assessment tools do create the opportunity for developers who are guided by medical evidence and do focus on providing effective experiences to set themselves apart from the vast majority of programs seemingly designed to cash in on an unmoderated market. However, such tools would have to find a way

of gaining attention of both developers and end-users to be of any use. They also need to be integrated with efforts to validate their actual efficacy.

4.3 Considerations for new validation tools and initiatives

Although we oppose the position that RCT designs take too much time by default, we realize that RCT designs are not optimized towards time. The aim for determining with certainty that results found are due to the interventions used, places all other considerations second. We also realize that RCTs and 'spinoff' methods are rooted in a specific academic culture favoring scientific explanation and knowledge innovation.

Since this report is *not* concerned with validation purely taking place within the confines of a university or other academic environment, the time investment only follows the amount of time required for the design, execution, and reporting of a specific game/app validation study. Since designs are readily available and can often be applied to multiple types of intervention, the first stage comes down to measurement development. These measurements can often be either reused from previous studies or tailored specifically for the objective of the intervention.

The second stage, then, is the study's execution. If the intended use time of a game/app is, for example, two weeks, then the execution should take no more than two weeks plus the time periods that are required for pre- and post-tests. Such a game/app can therefore be studied in a window of weeks or a few months, depending on how long effects would need to last.

The third stage is reporting. Since data-analysis depends on the complexity of the measures and design used, this stage could take up anything within a day's to a few weeks' time. Seen from start to finish, an experiment to validate the effects can be performed in under a month for very specific, micro-level interventions, or can take up to and over a year for comprehensive, longitudinal projects. Naturally, the kind of time investment is dependent on the intentions behind the intervention. For an evaluation of game/app merely as an *instrument* less time-consuming investments are possible because then comparisons with control groups are not necessary by default.

We currently observe, however, that game/app validation studies are often inextricably bound to academic careers (i.e. research careers at universities). Researchers write their dissertation on one or a few interventions, and their principal output (publications and conference presentations) are the main issue behind the slow turn-around of game/app validation studies. In this novel field, such a process is still advantageous to generalizable research. Finding best practices in design, establishing what kinds of methods allow researchers to find impacts, and building theoretical knowledge to predict the effects of future interventions are all part of the remit of the academic researcher.

However, validation of health games and apps could benefit from more diagnostic practical research, when e.g. only internal validity is at stake ("is there really an effect of this intervention?") as opposed to examining the game/app's construct validity ("what is the effective component of this game/app intervention?"), which in general is more time consuming.

It is not feasible, practical, or advantageous to study the close to 100,000 health-related apps that are currently available on the different online market places (Becker et al., 2014),

nor would it be wise to prohibit such interventions from being published without empirical evidence into their effects.

However, having a validated program should enable developers to stand out from the crowd, for example by being declarable through health insurance programs. Because this would be of interest to developers, the initiative and funding for validation should come from game studios (or the organization they are developing the program for). Insurance companies would in these instances be responsible for setting the standards for this research (design, execution, and reporting), preferably based on academic practices.

Looking at the Dutch landscape, this program has several advantages. First, it would allow insurance companies to only support those programs that have proven to be effective. Second, it would very likely thin out the amount of available apps, presenting less opportunity to unproven programs than to proven ones.

Naturally, protocols need to be designed to ensure that the research is rigorous and can indeed claim causal effects of interventions. Moreover, the procedure should not be equally stringent for each type of intervention (Boulos, Brewer, Karimkhani, Buller, & Dellavalle, 2014), and any future endorsements should reflect the kind of validation that was performed on a game/app and that was needed to meet the claims. Finally, we want to point out that an important consideration for the long run is to pay also attention to construct validity issues and using appropriate designs, because this allows for the isolation of effective components of a game/app, which can save time/efforts in designing and manufacturing new related games/apps.

5 Conclusion: A differentiated approach

5.1 A summary of the findings

In Chapter 2 we reported the outcomes of a substantial amount of meta-reviews concerning games in general and games/apps aimed at the healthcare industry. We concluded on the basis of this review that the general portrayal of the outcomes of serious games, also in the area of health, is positive. This conclusion is a tentative one, however, as issues with validity (related to the obtained data) and validation (related to the study design behind the data) often remain.

In Chapter 3 we distinguished four outcome variables where the use of games/app can be described: usability, playability, efficacy and side-effects. We concluded that it is useful to explicitly mention outcomes of the use of a specific game/app in terms of these 4 aspects.

Next we made a distinction concerning types of validity as this is a prerequisite in order to understand and build upon the issue of validity and the evaluation of games/apps in the health domain. Bearing this in mind, the most important validity types are:

a. Validity of a game/app as a measurement or assessment instrument.
   The central question is "does the game/app assess what we want it to assess?".
b. Internal validity of a game/app as treatment or intervention.
   The central question is, "is the game truly effective?"
c. Construct validity of a game/app as treatment or intervention.
   The central question is "which effective components of the game/app are responsible for the effect found?"

We mentioned a number of RCT designs and alternatives to RCT designs. Often these alternative strategies are next-best strategies and they can be useful when conditions for applying a RCT design are not or cannot be met. However, we also observed that it still has to be seen what the value is of the more advanced alternative RCT designs with respect to their time investment and needed statistical analysis expertise.

In Chapter 4 we first described methods and instruments to evaluate the usability and playability of apps/games. Next, we mentioned recent initiatives concerning (self) assessment of games/apps in terms of usability, playability, efficacy and validity. We conclude here that the exact way of doing this is still to be discussed, e.g. who is the curator of listed games/apps: insurance companies, governmental organizations (e.g. GGDs) or game studios themselves? This issue has yet to crystallize.

Our main conclusion is that the claims that are attached to a certain game/app determine the type of validity that should be checked, and at the same time the research design that is needed to examine those claims. This leads to a differentiated approach:

- The first question (or claim) is **to check whether the game/app is merely aimed at assessment of the game/app as an instrument** For instance, does this app measure anxiety management, or more generally how the player is gaining knowledge or training skills, or something related? If so, it would suffice to pay attention to content, construct and concurrent validity of the assessment.

Is the content of the game/app complete and is the intended goal measured well and does it correspond with alternative, established methods? While an experimental study will most likely need to be carried out, an RCT study design will not be required. In other words, often only measuring reference tests and computing correlations are in order, but not necessarily RCT study designs.

- **The second question (or claim) concerns whether the game/app as intervention is effective or not**. E.g. do users actually manage their anxiety with or thanks to the game/app effectively or not?

  This demands that the claim is examined whether a game group has improved more than a non-game, control group. In other words, given this claim, the internal validity has to be demonstrated, e.g. by using a RCT or next-best alternative design.

- The third question (or claim) is more detailed: here the question is not **simply whether the treatment/intervention is effective, but also an additional claim is examined: what is the effective component of the game?** Is it the gameplay, or is it simply the case that the monitoring activities are responsible for the results, irrespective of the game, etc.

  This issue involves the question of construct validity and often means breaking up the experiment in more detailed components or conditions, e.g. an extra control condition which, in this example, receives anxiety management information in a non-gaming way. It will be clear that proving the actual effective components needs complementary research.

It is important to stress that when it comes to games and apps, validation of efficacy needs to be complemented by validation of usability and playability. An app/game that is not usable/playable by the intended target audience will never reach its efficacy potential. We conclude that usability and playability do not by definition require RCTs or next-best alternative study designs, since usability and playability cannot be understood as effects of a cause (in this case the game or app). Usability and playability are properties of an app/game, not effects of it. This means that when a designer, developer or researcher *solely* wishes to validate the usability and/or playability of the app/game in question, we refer to the recommendation under the above first question/claim.

Having said that, when an RCT or next-best alternative study is planned to determine efficacy, the team can of course still choose to also incorporate measures for usability and playability. In any case, we envision that during the design phase of the project usability, playability and efficacy need to be parallel objectives for the team to focus on, while during the development phase it will prove more practical to first focus on usability and playability, and subsequently on efficacy of the game/app experience.

It is also important to stress that the validation of *side-effects* will certainly not always be relevant for health games and apps. We envision that it will often prove unfruitful and inefficient to focus on side-effects in the case of games/apps with objectives related to prevention that follow a learning approach (e.g. providing information/knowledge on the

adverse health effects of smoking). Having said that, we cannot exclude the relevance of side-effects for prevention-oriented health games/apps categorically.

Game developers, game studios, governmental organizations, researchers and other stakeholders such as insurance companies should be clear in the claims they want to make with a particular game/app, and the designs chosen should be able to support those claims.

5.2 Future research

This report offers sufficient basis for elaborating on the existing tools for health game/app validation reviewed in Chapter 3, as well as the frameworks they are based upon. The reviewed validation tools/efforts function well for (self-)assessment of health games/apps, but they do not yet explicitly cater to validation through appropriate empirical studies.

The differentiated approach we have explored for validation on this level needs to be developed and tested further on various games and apps for e-Health. Furthermore, it subsequently needs to be translated into practical tools that can complement (instead of replace) e.g. Graafland et al.'s (2014) checklist or the expert evaluation procedure of DSSH.

Thus, a first recommendation for future research concerns the further development and evaluation of the differentiated approach to validation. This should then be followed by the development and evaluation of a practical web-based toolset that aids a design/research team in performing an appropriate experimental study and analyzing its results in the validation process of a game/app. Integrating such a validation toolkit with other (self)assessment programs would help to normalize this typically research-focused practice of empirical effects-research as part of the quality assurance process of any game/app developer.

There is also a need for the development and evaluation of a broader risk assessment framework and toolset. After all, validation as analyzed in this report only covers one risk, i.e., the risk that a health game/app does not do what it was meant to do, or has unexpected, unacceptable side-effects. There are, of course, many other risks to consider, such as the risk of a breach of patient privacy or data security. While several tools reviewed in Chapter 4 take this broader risk assessment into account, we envision that further research into this topic will help elaborate and systematize the different types of risks involved in different types of health games/apps.

This report also offers sufficient basis for further professionalization of health game/app validation specifically in the Dutch market, with the goal of improving the quality of validation studies as well as speeding up these studies. Future research should be done examining the type of services that can be developed (e.g. validation packages, validation courses or training) to efficiently cater to the continuing validation needs of health game/app designers and developers as well as other stakeholders (e.g. insurance companies).

This line of research should also look into what organizational entities and business models would be suitable and preferable for these services, e.g. a separate entity of a knowledge institute (similar to the *Center for Research on User eXperience (CRUX)* at Utrecht University in the Netherlands), or a separate business altogether (similar to the company *Player Research* in the United Kingdom).

We expect that the above next research steps will greatly help the development of the health game/app market nationally and internationally. We strongly believe that further work into differentiated validation and risk assessment, their associated toolsets, and the subsequent professionalization of validation, will lead to a healthier and more competitive market for innovative e-Health games and apps.

References

Adams, S.A. (2010). Use of serious health games in health care: a review. *Studies in Health Technology and Informatics*, 157, 160-166.

Albrecht, U.V. (2013). Transparency of Health-Apps for Trust and Decision Making. *Journal of Medical Internet Research, 15*(12):e277

Baranowski, T., Buday, R., Thompson, D.I., & Baranowski, J. (2008). Playing for real. Video games and stories for health-related behavior change. *American Journal of Preventive Medicine, 34*(1), 74-82. http://doi.org/10.1016/j.amepre.2007.09.027

Becker, S., Miron-Shatz, T., Schumacher, N., Krocza, J., Diamantidis, C., & Albrecht, U.-V. (2014). mHealth 2.0: Experiences, possibilities, and perspectives. *JMIR mHealth and uHealth*, *2*(2), e24. http://doi.org/10.2196/mhealth.3328

Bonell, C.P., Hargreaves, J., Cousens, S., Ross, D., Hayes, R., Petticrew, M., & Kirkwood, B.R. (2011). Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions. *J Epidemiol Community Health, 65*(7), 582-7. http://doi.org/10.1136/jech.2008.082602

Boulos, M. N. K., Brewer, A. C., Karimkhani, C., Buller, D. B., & Dellavalle, R. P. (2014). Mobile medical and health apps: state of the art, concerns, regulatory control and certification. *Online Journal of Public Health Informatics*, *5*(3), 229. http://doi.org/10.5210/ojphi.v5i3.4814

Boyle, E.A., Connolly, T.M., Hainey, T., & Boyle, J.M. (2012). Engagement in digital entertainment games: a systematic review. *Computers in Human Behavior, 28*, 771-780.

Boyle, E.A., Hainey, T., Connolly, T.M. ...., & Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impact and outcomes of computer games and serious games. *Computers & Education, 94*, 178-192. http://doi.org/10.1016/j.compedu.2015.11.003

Brown, C.A. & Lilford, R.J. (2006). The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology, 6*(54). http://doi.org/10.1186/1471-2288-6-54

Cain, M. (2012). One company's experience: Blazing the trail with the first FDA-cleared medical imaging app. *Biomedical Instrumentation & Technology*, *46*(s2), 87–90. http://doi.org/10.2345/0899-8205-46.s2.87

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Boston, MS: Houghton Mifflin Company.

Cannon-Bowers, J.A., Bowers, C., & Procci, K. (2011). Using video games as educational tools in healthcare. In S. Tobias & J.D. Fletcher (Eds.), *Computer Games and Instruction*. Charlotte, NC: Information Age Publishing.

Cardullo, S., Gamberini, L., & Mapelli, D. (2015). Padua Rehabilitation Tool: A pilot study on patients with dementia. Paper presented *GALA2015 Conference*. Rome, Italy: GALA.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2015). Digital games, design, and learning: A systematic review and meta-analysis. *Review of educational research*, 0034654315582065. http://doi.org/10.3102/0034654315582065

Collins, L. M., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): New methods for more potent eHealth interventions. *American Journal of Preventive Medicine*, *32*(5 Suppl), S112–8. http://doi.org/10.1016/j.amepre.2007.01.022

Connolly, T.M., Boyle, E.A., MacArthur, E. Hainey, T., & Boyle, J.M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*, 661-686. http://doi.org/10.1016/j.compedu.2012.03.004

Cook, D.A., & Beckman, T.J. (2006). Currents concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*, 166.e7-166.e16. http://doi.org/10.1016/j.amjmed.2005.10.036

Cook, T.D., & Campbell, D.T. (1969). *Quasi-Experimentation. Design & Analysis Issues for Field Settings*. Chicago, IL: Rand McNally College Publishing Company.

DeSmet, A., van Rijckeghem, D., Compernolle, S., ... & De Bourdeaudhuij, I, (2014). A meta-analysis of serious digital games for healthy lifestyle promotion. *Preventive Medicine, 69*, 95-107. http://doi.org/10.1016/j.ypmed.2014.08.026

Desurvire, H., Caplan, M., & Toth, J.A. (2004). Using heuristics to evaluate the playability of games. *CHI 2004 Proceedings*. Vienna, Austria. http://doi.org/10.1145/985921.986102

Feinstein, A.H., & Cannon, H.M. (2002). Constructs of simulation evaluation. *Simulation & Gaming, 33*(4), 425-440. http://doi.org/10.1177/1046878102238606

Fitrianie, S., Griffioen-Both, F., Spruit, S., Lancee, J., & Beun, R.J. (2015). Automated dialogue generation for behavior intervention on mobile devices. *5th Int. Conference on Current and Future Trends of Information and Technologies in Healthcare. Procedia Computer Science, 63*, 236-243.

Gallagher, A.G., Ritter, E.M., & Satava, R.M. (2003). Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surgical Endoscopy And Other Interventional Techniques,* 17(10): 1525-1529. http://doi.org/10.1007/s00464-003-0035-4

Gamito, P., Oliveira, J.M, Lopes, P. ....., & Deus, A. (2014). Executive functioning in alcoholics following a mHealth cognitive stimulation program: Randomized controlled trial. *Journal of Medical Internet Research, 16*(4), e102. http://doi.org/10.2196/jmir.2923

Genetic Digital. (2013). When is an app classed as a medical device? Retrieved from http://www.geneticdigital.co.uk/2013/03/when-should-an-app-be-classed-as-a-device/

Ghanbarzadeh, R., Ghapanchi, A.H., Blumenstein, M., & Talaei-Khoei, A. (2014). A decade of research on the use of three-dimensional virtual worlds in health care: A systematic literature review. *Journal of Medical Internet Research, 16*(2): e47. http://doi.org/10.2196/jmir.3097

Graafland, M., Schraagen, J.M., & Schijven, M.P. (2012). Systematic review of serious games for medical education and surgical skills training. *British Journal of Surgery, 99*, 1322-1330. http://doi.org/10.1002/bjs.8819

Graafland, M., Dankbaar, M., Mert, A. ....., & Schijven, M. P. (2014). How to systematically assess serious games applied to health care. *Journal of Medical Internet Research Serious Games, 2*(2), e11. http://doi.org/10.2196/games.3825

Gray, W.D., & Salzman, M.C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*(3), 203-261.

Hamari, J., Koivisto, J., & Pakkanen, T. (2014). Do persuasive technologies persuade? – A review of empirical studies. In: Spagnolli, A., Chitattaro, L., & Gamberini, L. (Eds.), *Persuasive Technology*, LNCS8462, pp. 118-136. Bern, Switzerland: Springer International Publishing.

Ioannidis, J. P. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Jama*, *279*(4), 281–286. http://doi.org/10.1001/jama.279.4.281

JMIR (2012). Contribute to mHealth research and development of the new JMIR mHealth peer-review tool for mobile apps. http://www.jmir.org/announcement/view/77

Kato, P.M. (2010). Video games in health care: Closing the gap. *Review of General Psychology, 14*(2), 113-121. http://doi.org/10.1037/a0019441

Kato, P.M. (2012). Evaluating efficacy and validating games for health. *Games for Health Journal, 1*(1), 74-76. http://doi.org/10.1089/g4h.2012.1017

Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. In R.E. Ferdig (Ed.), *Handbook of Research on Effective Electronic Gaming in Education* (Vol.1, pp1-32). Hershey, PA: Information Science Reference.

Ke, F. (2015). Designing and integrating purposeful learning in game play: a systematic review. *Educational Technology Research and Development*. http://doi.org/10.1007/s11423-015-9418-1

Krebs, P., Prochaska, J.O., & Rossi (2010). A Meta-analysis of computer-tailored interventions for health behavior change. *Preventive Medicine, 51*, 214-221. http://doi.org/10.1016/j.ypmed.2010.06.004

Kueider, A.M., Parisi, J.M., Gross, A.I., & Rebok, G.W. (2012). Computerized cognitive training with older adults: A systematic review, *PLoS ONE, 7*(7) e40588. http://doi.org/10.1371/journal.pone.0040588

Kumar, S., Nilsen, W. J., Abernethy, A., Atienza, A., Patrick, K., Pavel, M., … Swendeman, D. (2013). Mobile health technology evaluation: The mHealth evidence workshop. *American Journal of Preventive Medicine*, *45*(2), 228–236. http://doi.org/10.1016/j.amepre.2013.03.017

Mayer, R.E. (2011) Multimedia learning and games. In S. Tobias & J.D. Fletcher (Eds.), *Computer Games and Instruction* (pp. 281-305). Charlotte, NC: Information Age Publishing.

Mayer, I., Bekebrede, G., Harteveld, C., ..., & Wenzler, I. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology, 45*(3), 502-527. http://doi.org/10.1111/bjet.12067

MHRA. (2014). Guidance Medical device stand-alone software including apps. Retrieved from https://www.gov.uk/government/publications/medical-devices-software-applications-apps/medical-device-stand-alone-software-including-apps

Mohr, D. C., Cheung, K., Schueller, S. M., Hendricks Brown, C., & Duan, N. (2013). Continuous evaluation of evolving behavioral intervention technologies. *American Journal of Preventive Medicine*, *45*(4), 517–23. http://doi.org/10.1016/j.amepre.2013.06.006

Moreno-Ger, P., Torrente, J., Hsieh, Y.G., & Lester, W.T. (2012). Usability testing for serious games: Making informed design decisions with user data. *Advances in Human-Computer Interaction*, Article ID 369637, 13 pages. http://doi.org/10.1155/2012/369637

NHS Choices. (2015). Health Apps Library. Retrieved from http://www.nhs.uk/pages/healthappslibrary.aspx

Olson, T., Procci, K., & Bowers, C. (2011). Serious games usability testing: How to ensure proper usability, playability, and effectiveness. In: A. Marcus (ed.), *Design, User Experience, and Usability. HCII 2011*. LNCS 6770 (pp. 625-634). Berlin: Springer-Verlag.

Pinelle, D., Wong, N., & Stach, T. (2008). Heuristic evaluation for games: Usability principles for video game design. *CHI 2008 Proceedings*. Florence, Italy. http://doi.org/10.1145/1357054.1357282

Portnoy, D.B., Scott-Sheldon, L.A., Johnson, B.T., & Carey, M.P. (2008). Computer-delivered interventions for health promotion and behavioral risk reduction: A meta-analysis of 75 randomized controlled trials, 1988-2007. *Preventive Medicine, 47*, 3-16. http://doi.org/10.1016/j.ypmed.2008.02.014

Primack, B.A., Carroll, M.V., McNamara, M.…, & Nayak, S. (2012). Role of video games in improving health-related outcomes: A systematic review. *American Journal of Preventive Medicine, 42*(6), 630-638. http://doi.org/10.1016/j.amepre.2012.02.023

Rahmani, E., & Boren, S.A. (2012). Videogames and health improvement: A literature review of randomized controlled trials. *Games for Health Journal: Research, Development, and Clinical Applications, 1*(5), 331-341. http://doi.org/10.1089/g4h.2012.0031

Renger, W.J., Veltkamp, R., & Schouten, B. (2015). *Towards a better risk analysis and validation of e-health applications in care and prevention*. Utrecht, the Netherlands: Growing Games.

Sitzmann, T. (2011). A meta-analytic examination of instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64*, 489-528. http://doi.org/10.1111/j.1744-6570.2011.01190.x

Soekarjo, M., & van Oostendorp, H. (2015). Measuring effectiveness of persuasive games using an informative control condition. *International Journal of Serious Games, 2*(2), 37-56. http://doi.org/10.17083/ijsg.v2i2.74

Subhi, Y., Bube, S.H., Bojsen, S.R. …. &, Konge, L. (2015) Expert involvement and adherence to medical evidence in medical phone apps: A systematic review. *Journal of Medical Internet Research mHealth and UHealth, 3(3),* e79, 13 pages. http://doi.org/10.2196/mhealth.4169

Van der Kooij, K., Hoogendoorn, E., Spijkerman, R., & Visch, V. (2015). Validation of games for behavioral change: Connecting the playful and serious. *International Journal of Serious Games, 2* (3), 63-75. http://doi.org/10.17083/ijsg.v2i3.75

Victora, C.G., Habicht, J-P., & Bryce, J. (2004). Evidence-based public health: Moving beyond randomized trials. *Public Health Matters, 94*(3), 400-5. http://doi.org/10.2105/AJPH.94.3.400

Vogel, J.J., Vogel, D.S., Cannon-Bowers, J., Bowers, C.A., & Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*, 229-243.

Warmelink, H., Valente, M., Van Tol, R., & Schravenhoff, R. (2015). Get it Right! Introducing a framework for integrating Validation in applied game design. Paper at the *GALA2015 Conference*. Rome, Italy: GALA.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., … Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*(8), 1359–1366. http://doi.org/10.2105/AJPH.2007.124446

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., & Wesslén, A. (2000). *Experimentation in Software Engineering*. Boston, MA/Dordrecht, NL: Kluwer Academic Publishers.

Wouters, P.J.M., van Nimwegen, C., van Oostendorp, H., & van der Spek, E.D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*, 249-265. http://doi.org/10.1037/a0031311

Wouters, P.J.M. & van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Computers & Education*, *60*, 412-425. http://doi.org/10.1016/j.compedu.2012.07.018

Yasini, M., & Marchand, G. (2015). Mobile health applications, in the absence of an authentic regulation, does the usability score correlate with a better medical reliability? *Studies in Health Technology and Informatics*, *216*, 127–131. http://doi.org/10.3233/978-1-61499-564-7-127

Appendix: Growing Games Validation Survey 2016

In the process of investigating the validation of e-health applications, we distributed an inventory survey among people who were active in this industry. This inventory survey was meant to gauge the attitudes of both developers and researchers towards the process of validating games and apps. After answering some questions on how involved they are with this practice, respondents gave their opinion on 13 attitude statements about validation. Each item was followed by an open answer section where explain their level of agreement with the statement. The answers to the closed questions are discussed here to give an overview of where respondents agree and where they disagree. Their remarks are included at the end of this appendix.

Twenty-eight people active in this field responded to the statements in the survey. Nineteen (68%) are currently developing e-health applications. Twenty-two (79%) are currently involved in researching e-health applications. The largest group of respondents (13, 46%) said they were primarily involved with e-health applications in their capacity as researchers, while eight (29%) were designers, first and foremost. Three respondents indicated they were involved with both research and development. Four respondents were involved in governmental or other organizations. On average, respondents have almost seven years (*M*: 6.6) of experience working with e-health applications. Their focus areas range from preventive health care, such as the importance of exercise for the elderly, to training programs, to help children cope with dyslexia for instance. While most of the respondents have participated in the validation of e-health applications, only one shared that the game she worked on is now covered by Dutch insurance companies. More information on this project can be found in the responses below.

Next, the responses to the statements relating to validation are discussed. Most of the respondents (86%) feel that there is a need for more validation of e-health applications in the current industry, reporting that more efforts to validate will make e-health applications more legitimate for the general public. Only 14% feels the community should focus on other things, rather than spend time on validation. At the same time, most respondents also say that the current ways in which these applications are validated are too expensive (54%), and that these practices take up too much time (64%). For both of these statements, there is a slight division between researchers and developers, where developers agree more strongly with this statement. Half of the respondents see a need for differentiated methods of validation, while 25% is undecided and the remaining 25% feels the application of a universal method would be best. Moreover, half the respondents feel that e-health applications that do not broach sensitive topics (do not have 'serious' intended effects) should be subject to less stringent validating tests than those that do (21% undecided, 29% disagree), with developers voting for more relaxed tests a little more than researchers. In any case, the majority of the respondents (82%) called for an exploration of other methods of validation that do not conform to the Randomized Controlled Trial model. When confronted with the statement that the validation of e-health applications should be left to academic researchers, 52% disagreed (19% undecided, 30% agreed).

On an organizational level, the majority of respondents (72%) indicated that legislature and regulations surrounding e-health applications should take the validation of such products into account, while only 4% disagreed. Similarly, most (72%) agree with the statement that insurance companies need to cover more e-health applications, whereby the vast majority

(82%) feels that greater focus on validation could help persuade these companies (researchers feel this more strongly than developers). The only statement the respondents really did not reach a consensus on was with regards to who would be responsible for funding validation efforts. Thirty-two percent agreed that developers should pay for it, while 21% disagreed (and 46% was undecided). Perhaps surprisingly, researchers were against developers paying for validation efforts slightly more than the developers themselves.

To summarize: the individuals who responded to our survey feel that not enough e-health applications are being properly validated. They feel that legislators could be more active in this regard as well, and that validation of such applications would serve to get more of them covered by health insurance plans. While it is certainly not unanimous, many respondents feel that the way in which the validation process works should be fine-tuned to fit the application, and that other methods to prove effectiveness should be explored. Although this is not a representative sample, the suggestions offered here do paint a picture of an e-health application community that is ready for a new way of working.

Responses to the statements

If you have ever been part of an effort to validate one or more e-health applications, could you tell us about your experience? 42

There is a need for (more) validation of e-health applications in the current industry. 44

Researchers: 44

Developers 45

Others 45

Validating e-health applications is necessary to make them more legitimate for the public. 46

Researchers 46

Developers 46

Others 46

Current methods to validate e-health applications are too expensive. 47

Researchers 47

Developers 48
Others 48
Current methods to validate e-health applications take too much time. 49
Researchers 49
Developers 49
Others 49
Different kinds of e-health applications should all be validated using the same methods. 50
Researchers 50
Developers 50
Others 51
An e-health application without 'serious' intended effects does not need to be validated as strictly as others. 52
Researchers 52
Developers 52
Others 52
We should focus on other things more than the validation of e-health applications. 53
Researchers 53
Developers 53
Others 53
Validating e-health applications can only be done by academics. 54
Researchers 54

Developers 54
Others 55
Regulations (and/or laws) need to take the validation of e-health applications into consideration. 56
Researchers 56
Developers 56
Others 56
Insurance companies need to cover (more) e-health applications 57
Researchers 57
Developers 57
Others 57
Insurance companies can be convinced to cover e-health applications if these are validated. 58
Researchers 58
Developers 58
Others 58
Developers/publishers of e-health applications need to pay for their validation. 59
Researchers 59
Developers 60
Others 60
Other methods of research besides RCTs need to be explored to validate e-health applications. 61

Researchers.....61

Developers 61

Others.....62

If you have ever been part of an effort to validate one or more e-health applications, could you tell us about your experience?

(5 years' experience): We have conducted several field experiments (feasibility and effectiveness) in two countries. This spring, we will conduct this kind of tests in a third country as well (also long-term tests, duration 6-8 weeks).

(5 years' experience): We are looking into different e-health solutions regarding promotion of mental health in school aged children and adolescents, in work well-being, health and safety as well as self-care competences regarding e.g. COPD.

(5 years' experience): I have been part of RCT study of active video games in PA promotion (of course other outcomes too). Long and time consuming process and it is a little bit waste of time to research only an intervention with games. Interventions (in patient populations) need to be more comprehensive. It is not useful (is what I think as a clinician) to test only the games. The games may be ONE part of an intervention. Then of course it is impossible to say what part of the intervention was effective. However, it is hardly ever [the case that] only one solution is enough for change in various or complex health problems.

(6 years' experience): At that time I was researcher at a development studio. However I was not the one undertaking the research. I was merely there to uphold the interests of the studio on the one hand and to help the researchers on the other. In the end I could say that everybody pretty much did what they wanted and did not listen that well to each other.

(10 years' experience): I am currently writing a meta-study on validation of health game features.

(5 years' experience): Difficult, costly, time consuming. Only way to let games land as serious business in healthcare.

(3 years' experience): We are testing a serious game if it works and what side effects it might have.

(3 years' experience): I received a subsidy of Kennisnet to validate the reading game Letterprins. However, as I am connected to Letterprins through intellectual property (Radboud University is the intellectual owner of Letterprins which I coordinated as inventor), I handed over the subsidy to independent researchers so they could validate the reading game without any conflicts of interest.

(4 years' experience): Long and arduous. Validation mostly used to increase the profile of the research organization, but not actually used as a tool in healthcare. Validation might not be the right way to make people actually use an app or game. Since professionals don't want

interference in their process or are afraid to use an innovation for fear of cost reduction which would slash their allocated budget if successful.

(4 years' experience): We developed our games as part of a PhD research track. Every step that was taken with regards to game design needed to be supported on an academic level, either with a literature review or a newly designed study or experiment. For us as game designers this went very slowly (it took four years in total), but in this way everything is validated. Everything that we did was carefully supported. If you want to know more, visit http://activecues.com/over-ons/voor-designers/ (*translated from Dutch*)

(7 years' experience): Previously I've been involved in several validation trials of our developed games. We have trialed different methods for in-house validations and verifications, but never a multi-center RCT. I'm currently involved in development and Validation of a game for rehabilitation (incl. RCT). This validation is performed in cooperation with an academic hospital.

(6 years' experience): I have been in many different validation processes, from traditional RCT's to designing new strategies with Universities and independent research companies. Traditional RCT's are far too slow for our industry and don't fit the development process of games. My experience is that many knowledge institutes lack the knowledge of both validation and game development to change things. Many companies lack the long term ambition to do it. We need both parties, in a long term commitment.

(20 years' experience): There is a will to validate, but the execution is not successful due to: - no researcher available during concept & development, only at the end of the project (too late) - no time planned by parties involved - no money invested - no resources available by parties involved.

(15 years' experience): RCT methods, with over more than 3 years before results and very difficult in include enough patients.

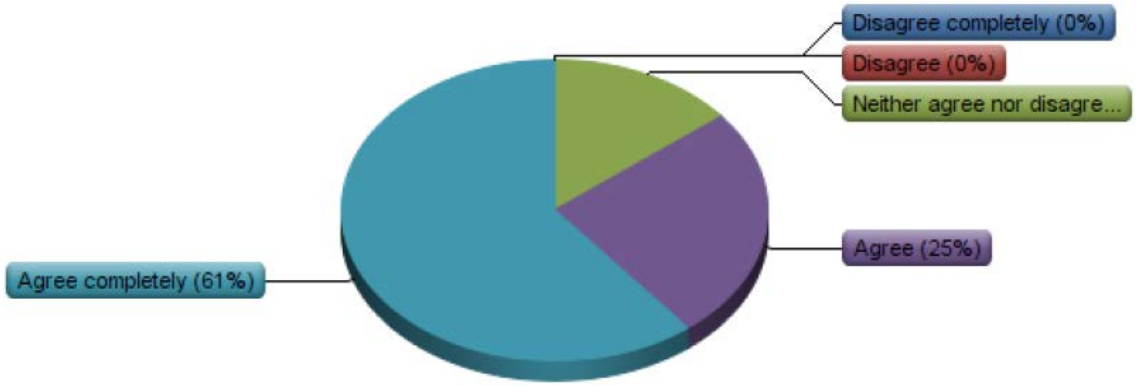
(17 years' experience): RCT way too expensive, way too long.

(17 years' experience): I have taken part in FDA validations.

(2 years' experience): It was a very hard and long process. Starting off, choosing the right method was already very difficult and it took a lot of time to find out the different ways. After my choice it took a lot of effort and time to do the validation and although it made the game more effective, it did made the game design process harder and took more time. Finally it was a challenge to be objective about the results... you really want your game to work!

(6 years' experience): It is a difficult and confusing experience. For starters the terminology used in the field as to what type of validation means what is far for conclusive. Performing full scale academic validation trials take a lot of time (depending on what type of validation is tested) and are costly and hard to get funding for. Also, these validation procedures so far are (too heavily) emphasizing parametrics over user experience.

There is a need for (more) validation of e-health applications in the current industry.



Researchers:

(5 years' experience): Yes, definitely. Before business we have to report exercise effect, safety issues etc. This is one of the key elements before starting any kind of global business activities.

(6 years' experience): This is a real Yes and No answer, but let me try to explain. Do we need full scale RCT's? No, generally I would say that this is not the case. However, if you want your game to pop up on a health insurance list, that is the way to go. Therefore we can say that we do need to link to the 'universal standard' of some branches the game industry is involved with. However, looking from another perspective we could say that games are something completely different than for instance; therapy, medicine or treatment. Of course, games are also developed to achieve these things, but they are fundamentally different. So in that way we could say that we need other standards that meet the needs of these new health phenomenon.

(10 years' experience): Many eHealth related applications are validated well, but many more (especially on wellbeing front) not at all.

(3 years' experience): There is need indeed, because currently these applications are not validated at all if they don't fall in to the category of a medical device which are regulated by EU. This need for validation is crucial especially related to applications meant for vulnerable groups such as children or people with physical or mental diseases. It would be good that even those applications that are meant even for health promotion of healthy people and contain some kind of health-related information would go through some kind of light validation process so that there would be some kind of control what the developers claim of their products and that the users would know if they can trust that the application contains the right information, for example. However, some applications that are meant for example to motivate someone in a fun way to do more exercises do not necessarily need the validation. However, if the developers want that this application would be recommended by a health professional to their patients, then the case is again different.

(3 years' experience): Only validated serious games can offer full benefit to society.

(4 years' experience): Validation is currently way too superficial.

Developers

(4 years' experience): This is not up to me. Will take too much time to formulate a weighted opinion about this. From a business perspective sales are all that matter. From a design perspective user experiences are all that matter (this doesn't translate as validation). From a health professional perspective validation is the only option available.

(6 years' experience): We need more validation and more different ways to do it. Many existing apps are not validated and to make apps that really make a difference the risks grow accordingly. Non-validated apps can only be implemented for non-risk subjects, these are mostly no stake subjects as well.

(20 years' experience): Too many well intended but poorly executed initiatives / products.

Others

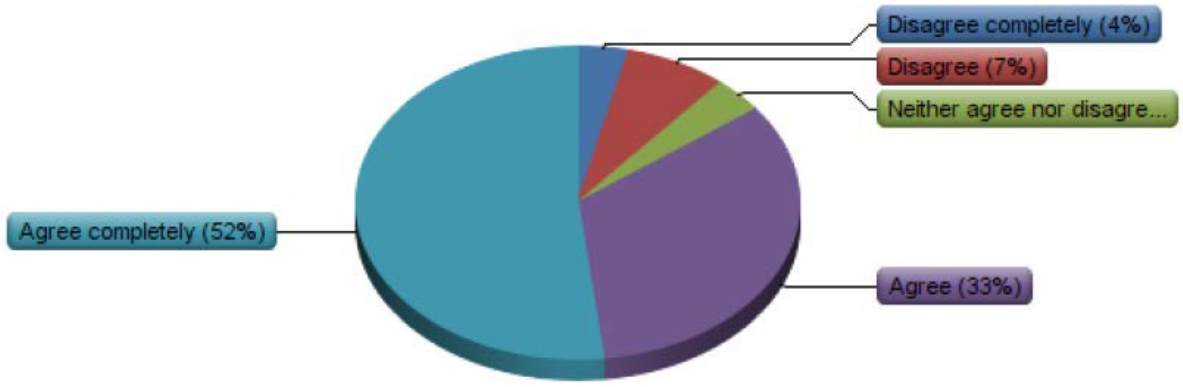
(17 years' experience): We need validation but a more practical form.

(17 years' experience): This is a broader question related to reliability; as eHealth apps become more numerous and multiform, doctors need to know just how much they can trust them (health impact, measurements, etc.).

(6 years' experience): There is the need for validation, however, it needs to be faster and more user oriented.

(4 years' experience): Faster validation and more (kinds of) validation.

Validating e-health applications is necessary to make them more legitimate for the public.



Researchers

(10 years' experience): Evidence based medicine - and health - should be what we strive for, instead of just profits by selling snake oil.

Developers

(4 years' experience): The public cares not one bit. Healthcare professionals care.

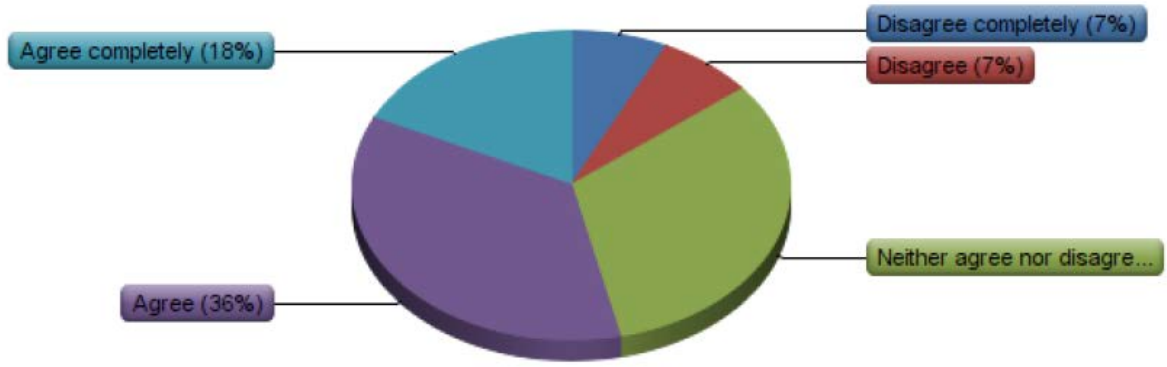
(6 years' experience): Not every app has to validated. It depends on the subjects. Validation does make the chance of successful implementation bigger.

(20 years' experience): Research & validation is necessary to develop successful products. It should be an integrated part of the development process. It's not just a "quality stamp" that is needed for marketing & sales. It's necessary for realizing meaningful effect for the end-user.

Others

(17 years' experience): Not for the public; consumers have different preferences and even though trust is an important factor, peer pressure and aesthetics can be of more importance if an application is acquired for personal use.

Current methods to validate e-health applications are too expensive.



Researchers

(5 years' experience): We as researchers can conduct quite cost-effective tests. Naturally tests in Singapore for example are quite expensive and time consuming including ethic approvals etc. but this is a part of the business and if you can do it you will be quite strong in this business.

(6 years' experience): This largely depends on what is demanded. You can also do small tests or little data calculations.

(3 years' experience): For example research requires always some money but not all applications need a long RCT study for the validation. However, the validation of a medical device is a very time consuming process.

(3 years' experience): Although our validation costs were higher than the developmental costs for Letterprins itself, it is worth it. Scientific research is expensive because it is elaborate and takes many man hours. However, in our case, university co-financed half of the project because it was so worthwhile both for the public as for pushing scientific boundaries. So, in my opinion, current methods are good and cannot be made less expensive. As a matter of fact, because we judged it to be unethical to work with the usual randomized controlled trials (half of the participants work with the game, other half doesn't and the progress of both groups is being compared), we conducted a kind of double RCT validation method that was twice as expensive: Our expectations of children with reading problems working with the game was so high, that we decided to pretest all children, then half of the children worked with the game Letterprins, other half didn't, all children were posttested, than we switched the groups and children who did not yet work with Letterprins now did, and the other half who already worked with Letterprins no longer did. We then tested the children again (retention test). This was far more elaborate and expensive than the usual validation research, however, we did show that working with Letterprins really makes children progress in their reading scores ánd it was ethical not to leave out any children from working with Letterprins.

Developers

(4 years' experience): There is only one method that is categorically appointed as the go to validation method: [RCT]. Which is an archaic way to look at validation.

(4 years' experience): It is time-consuming in any case (*translated from Dutch*)

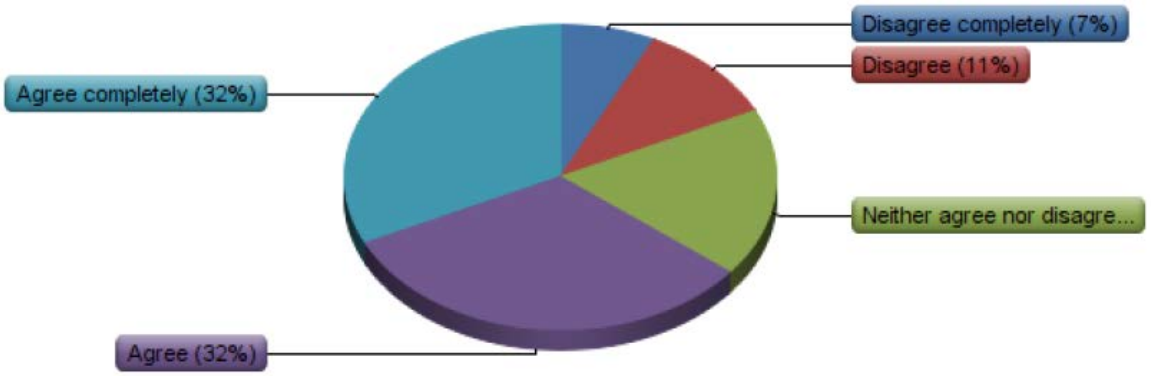
(6 years' experience): Many methods are not. There's too much focus on RCT.

(20 years' experience): There is a huge gap between the processes of e-health developers, researchers and health organizations. This must be organized more efficiently. All parties should respect and adapt these processes in order to work more time & cost efficiently.

Others

(17 years' experience): There are different degrees of validation. In higher tiers cost is a factor but not on lower tiers (such as, when the application is intended as informative and not related to actual care). It is actually a question of market strategy.

Current methods to validate e-health applications take too much time.



Researchers

(5 years' experience): it's understandable that we need long-term tests to identify effectiveness. If we compare this one with other scientific tests it's not too much at all (e.g. tests with new drugs)

(6 years' experience): I think that overall they are not suited for the fast changing pace that games have incorporated (multiple versions, monthly updates, bug fixes, etc.)

(10 years' experience): If validation is integrated in the plan from the beginning, it is not that hard to plan for, really.

(3 years' experience): Good validation research simply takes time.

(4 years' experience): In my experience they do not take enough time.

Developers

(7 years' experience): For a clinical trial you need 3 years. By then all the technology has become obsolete and you are back on square one. (*translated from Dutch*)

(4 years' experience): To come to successful design a game needs to be refined in small iterative cycles, this is completely contrary to the slow process of RCT'ing a game. Needs to be faster, low threshold to start validating, and patient oriented.

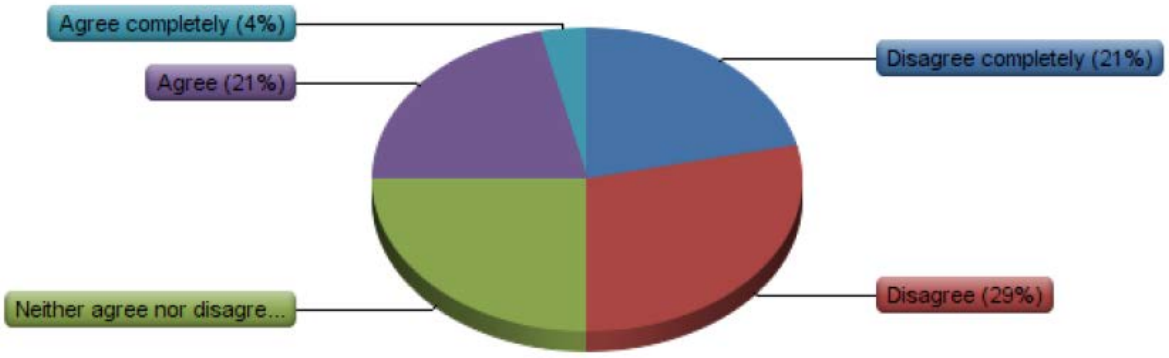
(6 years' experience): It can be done much quicker. Universities are slow. Researchers don't have to be.

(20 years' experience): There is a huge gap between the processes of e-health developers, researchers and health organizations. This must be organized more efficiently. All parties should respect and adapt these processes in order to work more time & cost efficiently.

Others

(17 years' experience): Again, depends on the degree.

Different kinds of e-health applications should all be validated using the same methods.



Researchers

(5 years' experience): It's up to the research approach. For example when we were first time in Japan we were utilizing Kansei Engineering research philosophy to learn more about how tests are conducted in Japan. Naturally to be able to compare results between Finland and Japan for example the same methods have to be used. So many kinds of validations are needed. One interesting research topic is cultural differences.

(6 years' experience): Not possible. Some mutual criteria might be useful but research with same methods is not always possible.

(10 years' experience): EHR requires clearly different methods than a health game.

(3 years' experience): It depends on which group it is intended for and for what purposes.

(1 year's experience): The methods should be the same in principle. We need to look critically at the kind of apps involved. I think game designers that do not want their games to be prescribed by doctors shouldn't want their games to be validated. (*translated from Dutch*)

(3 years' experience): Of course it would be nice if e-health applications can be compared, however, I think that this is not always possible. The effect sizes however can be compared (Cohen's D or partial eta square) to get an impression.

(2 years' experience): There could be some common parts, but e.g. physical and cognitive applications are quite different.

Developers

(4 years' experience): Every case has a different context. Every use is specific. Every validation needs to be context specific.

(4 years' experience): This seems the most ridiculous thing. An e-health app for patients with COPD is wholly different than one that helps children with anxiety disorders. The validation should be performed in different ways because of that. (*translated from Dutch*)

(20 years' experience): That really depends on the context of the application and environment it is used for.

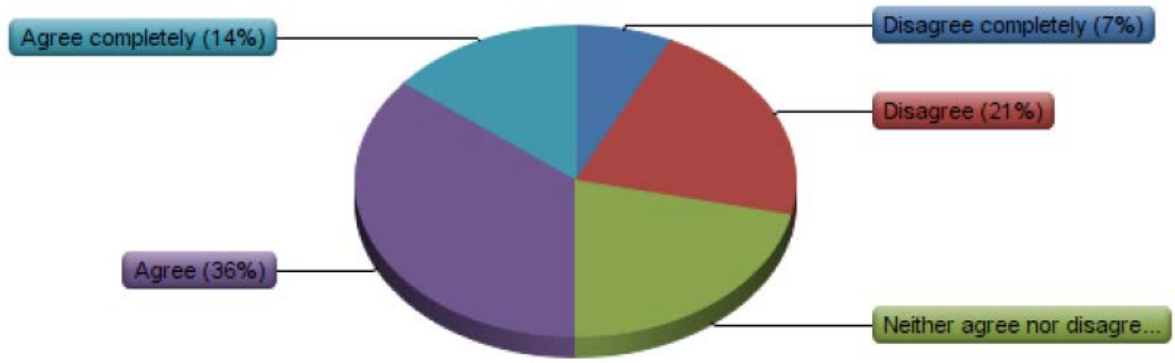
Others

(15 years' experience): Depends on the purpose, the expected outcome and target group.

(17 years' experience): Low impact tools like lifestyle coaches demand a different approach than for example physical therapy.

(17 years' experience): Methods should vary depending on the tier. If the applications are part of care, they should be validated *and* certified. This is a trust issue.

An e-health application without 'serious' intended effects does not need to be validated as strictly as others.



Researchers

(5 years' experience): We have to validate everything as seriously as possible. It's understandable that if the operations are more or less in wellbeing rather than in medical treatment requirements are totally different for the validation.

(16 years' experience): Rather than talking about strictness, I would recommend the term different.

(10 years' experience): Well, either it is a health application or not - and if it is, there needs to be proof that it actually positively affects health.

(4 years' experience): Define serious? Do we really need the eHealth homeopathy?

Developers

(6 years' experience): It is not something to generalize.

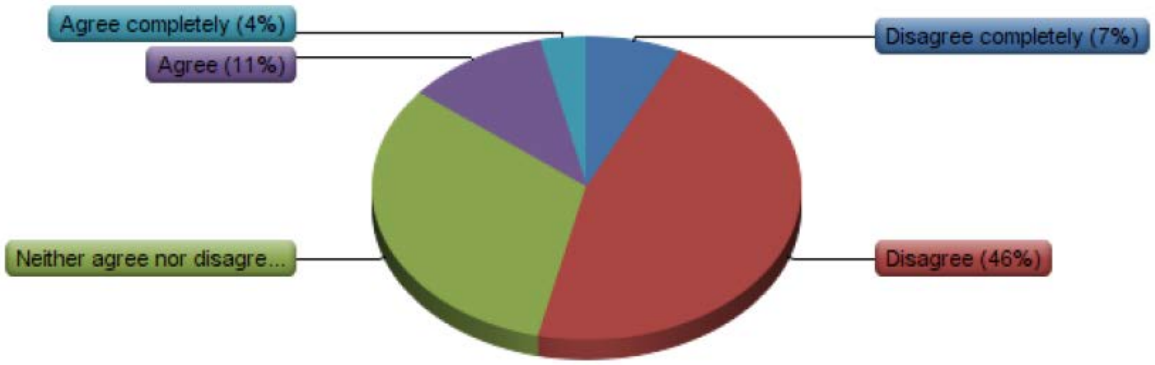
(20 years' experience): I suppose the same rules apply for e-health applications as other (serious or not serious) health interventions. Use those. Don't invent them again just because it is digital.

Others

(17 years' experience): Normal market place rules are sufficient. Consumers tend to report if applications are... less-than-informative.

(2 years' experience): Not as strictly... but it definitely needs to be validated in one way or the other to disprove negative effects.

We should focus on other things more than the validation of e-health applications.



Researchers

(5 years' experience): We have to focus on validation but many other aspects as well. Concept design, business ecosystems, internet of things etc.

(6 years' experience): This depends so greatly from company to company.

(10 years' experience): Privacy, getting people to use health applications etc. are of course also important, but so is validation.

(4 years' experience): There are other areas in eHealth that also need enhancement, but that is not an excuse for not validating them.

Developers

(1 year's experience): Quality over quantity.

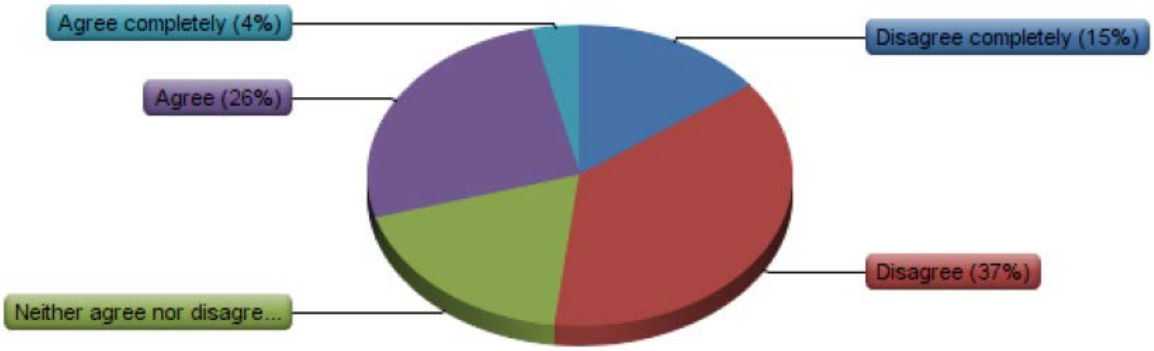
(4 years' experience): It depends on the app. If you claim you can cure cancer with your app, you need to be sure you can validate that. But if you help people to have more pleasant dreams, maybe validation is a little less important. (*translated from Dutch*)

(20 years' experience): Research & validation is one of the things a developer and the client should focus on.

Others

(17 years' experience): Validation *and* verification. There should be something like "continua certified" in the EU.

Validating e-health applications can only be done by academics.



Researchers

(5 years' experience): We can do it but the business perspective is also important!

(5 years' experience): Academics may only do research that is needed for validation statements. But the criteria (made by academics) can be used by others too for example health care professionals.

(16 years' experience): Parts of the validation process demand research. Somebody in the team has to be able to do that.

(6 years' experience): Creating the methods for validation should be done by academics in order to let more people actually do the validation.

(10 years' experience): Professional people can of course do this as well - if not often better! They can, after all, concentrate on the actual work without the academic pressures added.

(5 years' experience): Serious effects: yes. Less serious effects: no. However, requires strict and impartial evaluation. Not by developers themselves.

(3 years' experience): A health professional could do it as well (or maybe even someone else after specific education and good framework that guides the validation).

(1 year's experience): It could be done by patient interest groups, but for scientific validation science is certainly necessary. (*translated from Dutch*)

(3 years' experience): Only academics have the experience to validate applications and judge the results properly, using literature as a reference. However, I am convinced that academics in this area should cooperate with the practitioners. This is exactly what we have done bot in developing Letterprins and in validating Letterprins. A multidisciplinary team was involved in the validation of Letterprins.

Developers

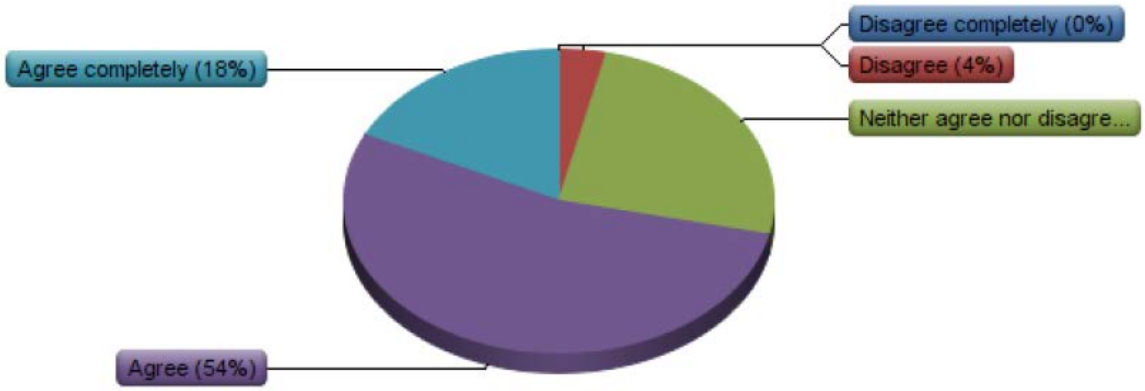
(4 years' experience): No, but they are important. An academically designed validation test is important. (*translated from Dutch*)

Others

(17 years' experience): End-user perspective and first-hand experience on the use are of importance. Some factors could be validated by a consumer panel (or similar) while others (health impact) should be validated by domain experts.

(2 years' experience): At the moment it does. We need standardized effective models/methods.

Regulations (and/or laws) need to take the validation of e-health applications into consideration.



Researchers

(5 years' experience): It's depending on authorities. For example between Finnish and Asian regulations we can find a lot of differences.

(4 years' experience): Regulation/law would mean that there is enforcement of the rules.

Developers

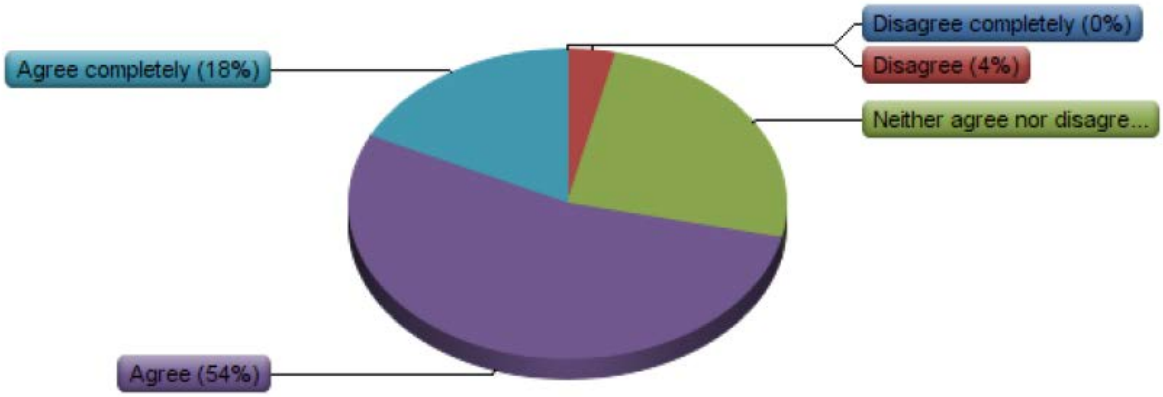
(20 years' experience): I suppose the same rules apply for e-health applications as other (serious or not serious) health interventions. Use those. Don't invent them again just because it is digital.

Others

(17 years' experience): Depending on the application, its use, and add-ons some (rarely all) of the following directives apply: 93/42/EEC; 90/385/EEC; 98/79/EC.

(6 years' experience): Maybe it should be legislated but it can't be the regular uptight Dutch way of doing things.

Insurance companies need to cover (more) e-health applications.



Researchers

(5 years' experience): This is very interesting question. I believe that insurance companies will be in a big role in the e-health ecosystems in the future.

(6 years' experience): They want to, but jeesh, they hold their standard tight and firmly. And why wouldn't they? They got the power!

Developers

(7 years' experience): Only the insurance companies will benefit from e-health, and they hardly ever take part in projects. (*translated from Dutch*)

(6 years' experience): Not in the current state, most of them don't have any effect.

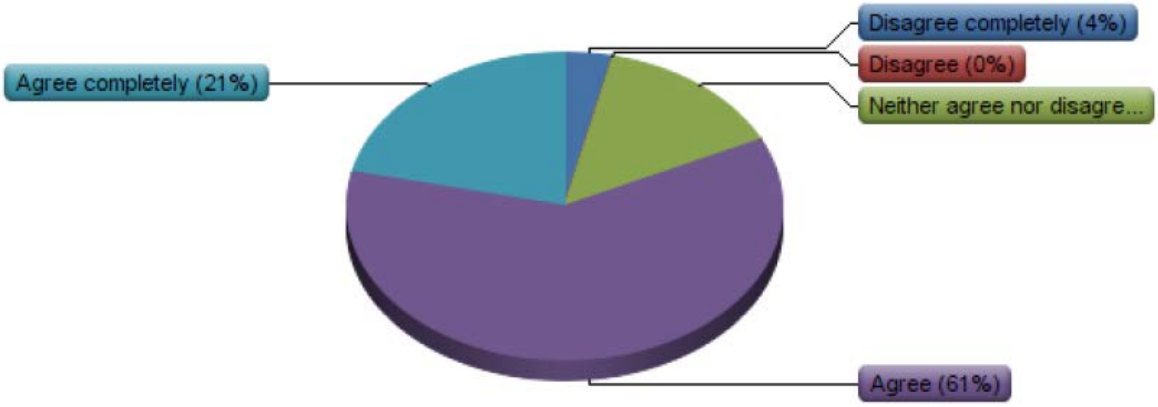
(20 years' experience): I suppose the same rules apply for e-health applications as other (serious or not serious) health interventions. Use those. Don't invent them again just because it is digital.

Others

(17 years' experience): The trick is how to translate this demand into business (and consumer incentives). "Why to bother".

(2 years' experience): When the applications are of better quality.

Insurance companies can be convinced to cover e-health applications if these are validated.



Researchers

(5 years' experience): Yes, we have already seen that they are interested in.

(6 years' experience): Talked enough with them, they are ready, as long as we validate that stuff in the way they like!

(10 years' experience): If it can be shown that people who use them cause less costs, sure. Then, how to verify that the people who claim to use them actually do. Especially without breaching privacy of the users.

Developers

(7 years' experience): To validate them you already need the insurance companies. If they are already validated you do not need them anymore. This needs to happen at an earlier point in the process. (*translated from Dutch*)

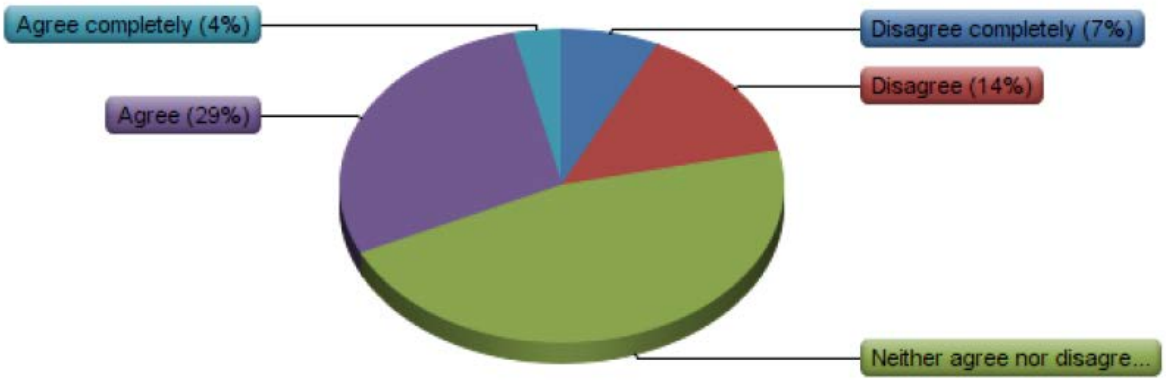
(4 years' experience): Change 'can' into 'should'.

(20 years' experience): I suppose the same rules apply for e-health applications as other (serious or not serious) health interventions. Use those. Don't invent them again just because it is digital.

Others

(17 years' experience): If the health impact is also related to lowering costs of care (home-based measurements vs. measurements at a health facility).

Developers/publishers of e-health applications need to pay for their validation.



Researchers

(5 years' experience): Naturally we have to pay but we can always seek funding instruments.

(16 years' experience): Work costs.

(6 years' experience): This depends greatly on what kind of project developers undertake.

(10 years' experience): Could (and often should) of course be whoever ordered the application.

(3 years' experience): Not at least for a lot of money.

(3 years' experience): This is absolutely the trickiest question so far in this survey... If the developers of e-health applications pay for its validation, conflicts of interest may arise and as is seen more often in validation research, the researchers are sometimes 'pushed' to find certain outcomes. If the developers pay for validation, independence of research must be guaranteed! In our case, which is to be favored, the validation was paid for by an independent financer and even co-financed by university. Of course, this can only be done if the serious game is só worthwhile that all parties see its surplus value, both for the public and for science itself. It leaves the scientists in an independent position without any conflicts of interest, yet as the scientists themselves are triggered by curiosity about WHY and FOR WHICH SUBGROUPS the serious game is working well, they will be urged to search for more funding themselves. This is a long-term cooperation (public-private collaboration) that works really well, it just asks parties to become involved in the concept of the serious game and that parties really see how the serious game can change existing practice in something better. This is exactly what is happening: the major validation research is published in several journal and 2(!) new validation research projects are being conducted at the moment digging further into Letterprins' validation, and one more subsidy application was submitted last month.

(4 years' experience): Somehow the cost will go to devs and publishers anyway, and from them to clients.

Developers

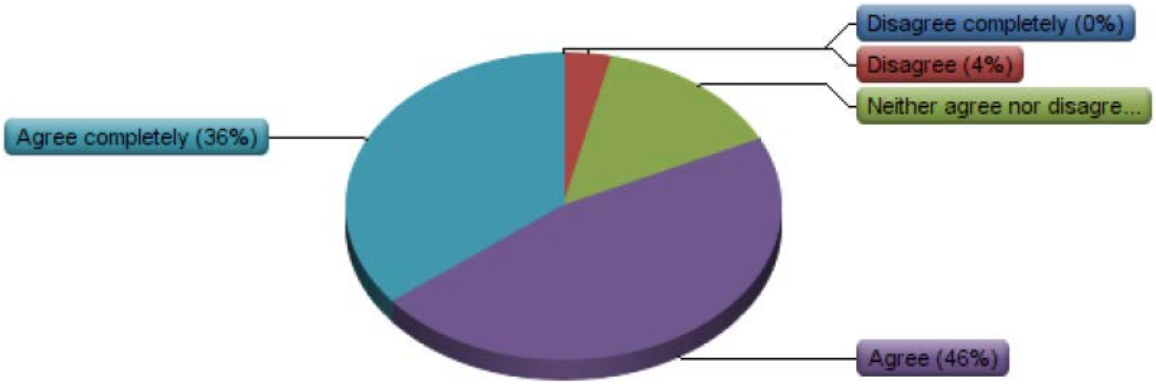
(20 years' experience): That depends on the business case and the amount of partners involved.

Others

(17 years' experience): It is a common practice in diagnostics.

(2 years' experience): They can validate them themselves, although there needs to be an objective reviewer.

Other methods of research besides RCTs need to be explored to validate e-health applications.



Researchers

(5 years' experience): Everything has to be explored.

(5 years' experience): All needed for different purposes.

(5 years' experience): RCT is only useful for certain research questions.

(3 years' experience): Evaluating the development process (who have been involved, what information is included etc.) is equally important. Also other lighter research methods than RCTs can be beneficial as well, such as Delphi panels, pilot studies, quasi-experimental trials and even qualitative studies.

(1 year's experience): Haven't these other methods already been explored? As stated previously, not every app needs to be fully validated, but the issue is that game designers shouldn't want to in the first place. (*translated from Dutch*)

(3 years' experience): As a first step, I think RCT is absolutely necessary to validate a certain application. However, I favor MORE validation as is done for Letterprins. One of the follow-up validation projects digs into the implementation of Letterprins. The RCT already showed us that it does work in a controlled setting, however, what are its benefits when it is used in a real-time setting of usual practice? I would like to refer to the standard work of how health innovations are to be implemented into clinical settings: this is usually an entire cycle of implementation. See Grol, R. & Wensing, M. (2011). Implementatie. Effectieve verbetering van de patiëntenzorg. Reed Business Education. These lessons for health innovation hold for e-health apps and games as well to my opinion and they should not be treated differently from regular health innovations.

(4 years' experience): RCTs maybe the "gold standard", but are they truly applicable on all kinds of validations?

Developers

(4 years' experience): This is most important.

(20 years' experience): I suppose the same rules apply for e-health applications as other (serious or not serious) health interventions. Use those. Don't invent them again just because it is digital.

Others

(17 years' experience): I think different tiers need different methods. Some lighter, some more demanding. In some cases (lowest tier) consumer rating and feedback is enough.

(6 years' experience): RCTs are neither feasible nor necessary as the ONLY form of validation. Better yet, every application should start at lower levels of validation; face validation, ecological validation and playtesting for the 'fun-factor' should already be considered valid forms of validation.

This report was drafted as part of the Growing Games project. Growing Games is a long-term stimulus programme (2013-2016) to promote the sustainable growth of the Dutch applied games sector. See www.growinggames.nl.

Authors:

UU	Herre van Oostendorp
HKU	Harald Warmelink
EUR	Ruud Jacobs

Editorial coordination:

iZovator/Growing Games	Doret Brandjes
DGA/CLICKNLgames/Growing Games	Irmgard Noordhoek
DGG/Growing Games	Marilla Valente, Christel van Grinsven
UCREATE	Karin Alfenaar

Other consortium members:

G4H	Tini Elemans, Sandra van Rijswijk
TNO/Growing Games	Christiaan van den Berg, Luuk Engbers, Esther Oprins
Gainplay	Teun Aalbers
TUDelft	Asli Boru
EBU	Jelle van der Weijden
UMCG	Monique Taverne, Kiki Spanjers
Ranj+TUDelft	Michael Bas
HKU	Willem-Jan Renger
UU	Remco Veltkamp
HvA	Ben Schouten
IJsfontein	Evert Hoogendoorn
-	Riëtte Meijer