2. Evaluating the value of technology

2.1 Measurement quality

2.1.1 Why evaluate?

Evaluation of technology is not typically considered as one of the more glamorous aspects of working in sport, but it is certainly one of the most crucial. A multitude of reasons exist with respect to why all technology should be evaluated (and sometimes require ongoing evaluation) prior to implementation in a sport:

- *Quality control:* Questions relating to quality include the following: Does the technology achieve what it purports to do – and does it do so consistently? Are there meaningful differences between different versions of the technology? How resilient is the technology when utilised in the field – will it maintain its quality in the long term? Quality is inextricably linked to validity and reliability, both of which concepts are discussed in greater detail in Theme 2.1.2 and 2.1.3 below.
- *Capability:* Questions relating to capability include the following: What use does the technology have within the organisation, other than for its intended one? Could it be used for other purposes, or combined with existing devices? Can it be built on and expanded upon for new initiatives and projects into the future?
- *Safety:* It is self-evident that technology should be deemed safe for use before being promoted or mandated for use in sports training or competition. However, sometimes achieving this is not as easy as it seems. Conducting a valid safety experiment in the same dynamic, complex environments that are typically experienced in competition environments is almost a diametrically opposed activity. It can also be a slow and even an expensive process. Further, in some cases, it is the most extreme, unlikely event that needs to be tested for a technology in order to truly understand its safety; these events are often hard to replicate in a testing scenario.
- *Benchmarking:* If other similar technologies or manufacturers of the same technology exist, how does the technology compare? How does a new release of a technology compare against a previous version?
- *Value:* You have to consider the following questions: Is it worth the cost or does it deliver a return on investment? Will the outlay deliver long and short term benefits? What is the cost of not utilising the technology?

A plethora of work exists within the sport literature by which assessments or testing instruments can be assessed with respect to their quality and appropriateness for use in the field (see Currell & Jeukendrup, 2008; Robertson, et al., 2014). Many of these works propose frameworks with groundings in measurement theory or psychometrics; some are readily transferable to the evaluation of technology. One of these (Robertson, et al., 2017) proposes that the three categories can be considered in order to define the measurement characteristics (validity, reliability & responsiveness, and sensitivity). The Feasibility properties of the technology should also be evaluated, in order to determine its applicability in the field. These properties are discussed throughout the remainder of this module.

2.1.2. Validity

Most individuals with even minimal scientific education will have at least some understanding of the concept of validity. What is less well-known is that multiple types of validity exist. A number of these are relevant for consideration when determining the quality of a technology:

- *Concurrent:* This type of validity seeks to determine the relationship of the output of a given technology with other similar measures that are obtained at the same time (Streiner, et al., 2015). For instance, if a new method of measuring VO2 maximum is compared with an established method on the same participant, the difference it records from this standard can be taken as a representation of its concurrent validity (when the established method is recognised as a 'gold standard', concurrent validity can be considered as criterion validity, instead). In the above example, a VO2 maximum recording of 65 mL/(kg.min) vs. an 'actual' recorded value of 64 mL/(kg.min) would clearly represent a minor overestimation. However, these examples are seldom this simple in reality, as further context on how the data was collected is normally required.

  The best way to determine concurrent validity of a technology is to assess it directly in a competition environment (sometimes referred to as 'ecological validity'). However, this is often not done at least initially, due to safety or feasibility reasons. Instead, test settings are typically employed, either in a laboratory or controlled environment.

  Another complicating aspect of this process is the understanding that no method exists by which to establish concurrent validity. This is because results may vary depending on the format of the data that is used to make the comparison, along with the evaluation metric used. Consider an example whereby an electronic performance tracking systems (EPTS) operating at 10 Hz is measuring a player's

displacement throughout a football match. At a basic level, a simple distance comparison with an established method could provide some insights into the system's concurrent validity. If the established method recorded the player as covering 10,000 m and the EPTS system recorded 10,100 m, it could be stated that it overestimates tracking distance by 100 m for a typical match, or 1%. However, is the equation really that simple? As team sport competition is dynamic and variable from match to match, more information is required in order to gain a more complete representation of validity. For the example above, in order to assess ecological validity (at least partially), it would be important that the testing conditions experienced represented those similar to those in which the technology would be used in competition (see Figure 2.1.2.a for a hypothetical example).

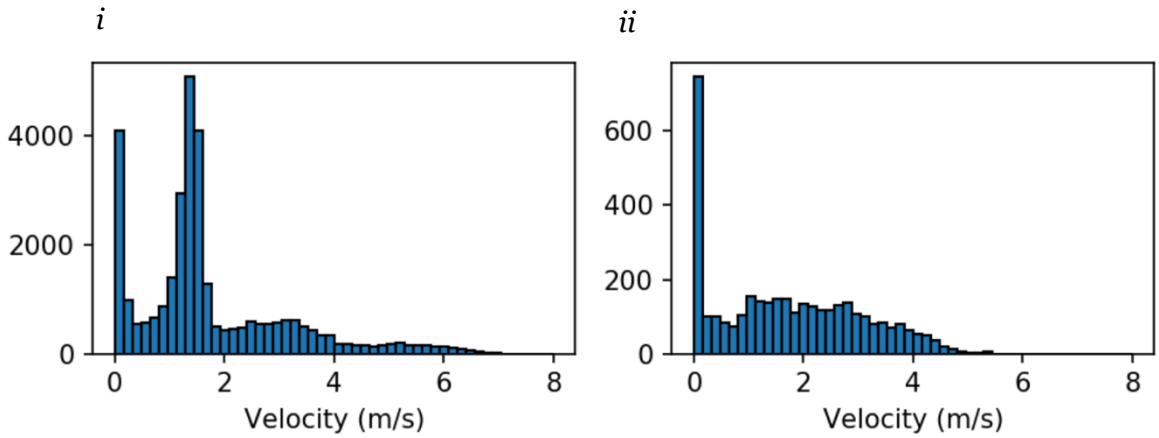

Figure 2.1.2.a. Distribution of velocity by a footballer in i) in a match setting and ii) testing the concurrent validity of an EPTS. Data was collected at 10 Hz, and the y axis represents the number of samples in the corresponding x axis velocity bin.

.

In the visual, the distribution of velocity recorded in competition is displayed on the left (i), with the testing session on the right (ii). It can be seen that although the proportion of time spent at different velocities is similar in both settings, minor differences exist. Specifically, there is a greater proportion of time spent walking (i.e., ~2 m/s) in competition, whereas in the test there is more time comparatively standing still (~0 m/s). If the samples were identical, then 'perfect' ecological validity could be said to be present. Because they are not, if a particular system is more or less accurate at one velocity compared to another, these differences may result in an over or underinflated perspective of the technology's performance to ensue. Figure 2.1.2.b illustrates differences in measurement between two systems by showing the proportion of samples that were in agreement, or under- and over-estimated (see figure caption for further information). A density of samples to the right of the zero

line on the *x* axis (as is shown in the example) indicates a higher number of overestimated data points; to the left of the zero, line indicates underestimation. Figure 2.1.2.c reports the same data in an alternative way, in which the differences between the test system and the established measure can be communicated.

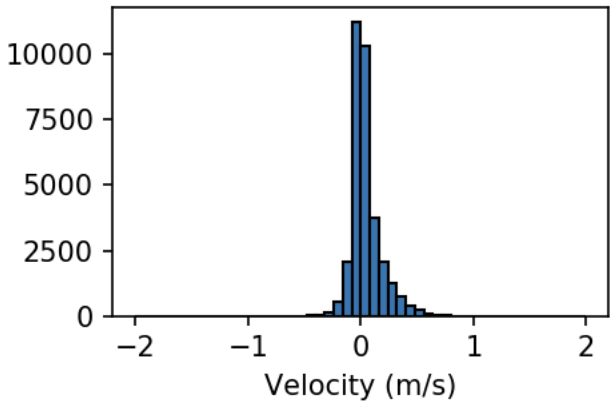

Figure 2.1.2.b. Distribution of differences in velocity calculations by an exemplar EPTS system against a gold standard. The y axis represents the number of samples and, thus, a higher density centred around the zero mark. The x axis indicates stronger agreement between the two systems.

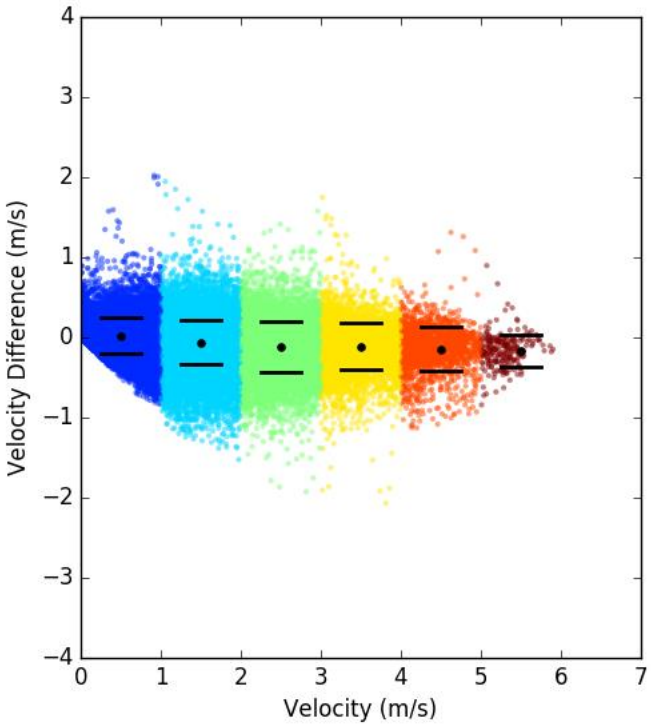

Figure 2.1.2.c. Distribution of differences between established method of measuring velocity (x axis) and tested method (y axis). Each point represents 0,1 s (1 sample).

Colour has been used to arbitrarily band the data into 1 m/s increments to aid any visual comparisons.

On the *x* axis, each point represents a sample and a corresponding instantaneous velocity; the point on the *y* axis represents the extent to which the tested technology recorded a higher or lower reading. If the values exhibit a greater congregation above the horizontal zero line, then there is a systematic overestimation by the tested system against the established method, and vice versa. Thus, this graphic displays the exact same data from figure 2.1.2.b, but using a different visualisation technique. However, here the effect of velocity on the differences between systems can also be investigated. For instance, a greater dispersion of recorded values away from the horizontal zero line in both directions at higher velocities compared to lower velocities (sometimes referred to as 'heteroscedasticity') means that the systems are in lower agreement when the athlete is running as opposed to when they are walking or jogging. Another way in which the same data can be visualised with the end user in mind is by examining the data based on its spatial properties (figure 2.1.2.d).

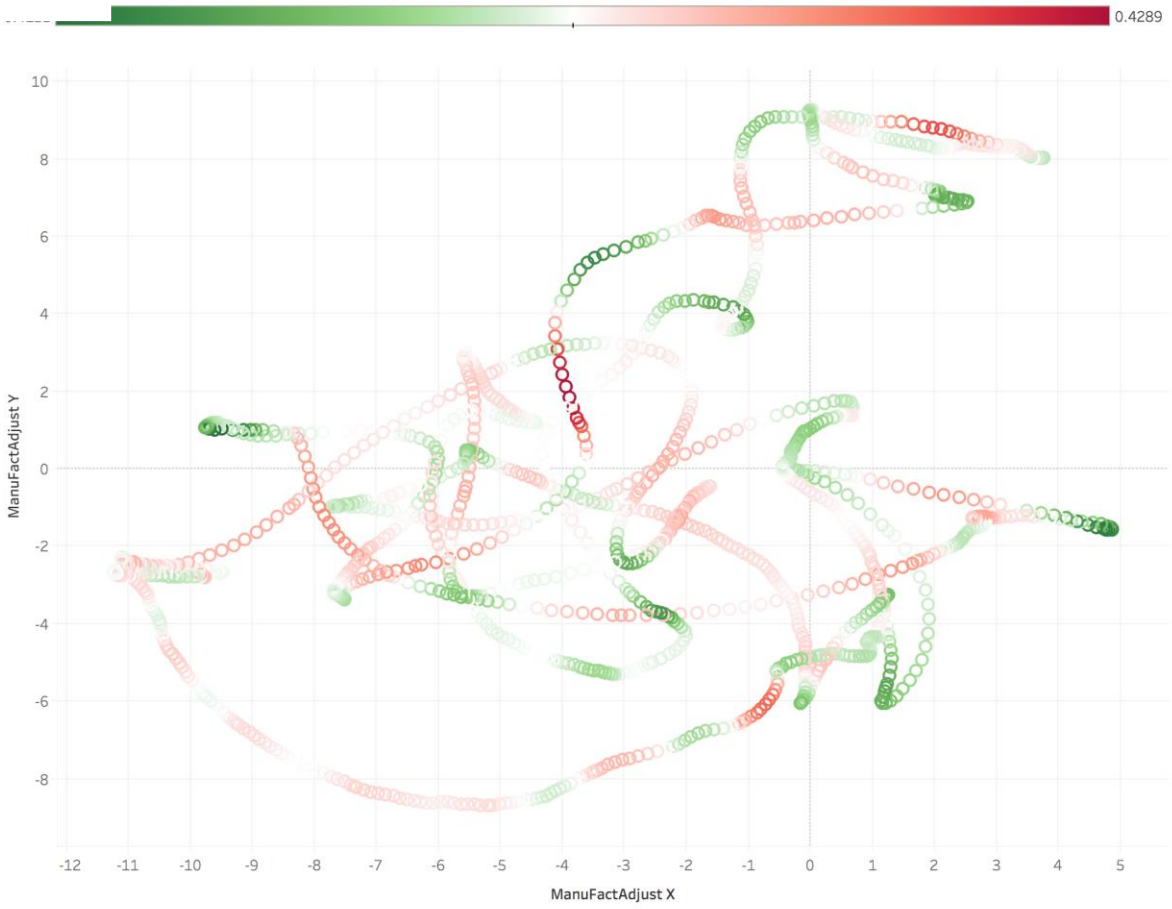

Figure 2.1.2.d. Representation of over- or under-estimation of an EPTS system against an established method. The trace represents the XY location of a player. Darker red indicates a section of underestimation, whereas darker green indicates a section of overestimation.

In this example, the *XY* location of the player is represented, with the difference in recorded velocity by the tested system to the established method visualised using colour. In this example, darker red indicates periods of the trace where the tested system reports higher velocity than the established method, with green representing an underestimation. Combined with video, this serves to provide greater insights into the types of movements that lead to the system showing lower agreement with the existing method.

The figure above represents just a few of the possible ways by which the validity of the system can be evaluated through various contexts and visualisations. The possibilities by which these comparisons can be made are almost endless. As should

be clear, however, if a single measure or visualisation of agreement or validity for the system is considered, then, this is likely to be inappropriate for use in the field, irrespective of the type of technology utilised. Thus, the main takeaway message is that seldom will a single number provide a good representation of a technology's validity and that a variety of metrics, analyses and visualisations is required in order to obtain a more complete understanding.

- *Content:* This form of validity is more conceptual than concurrent, in that it refers to how much of a given construct the technology measures or develops an understanding of content (Robertson, et al., 2014). Using the previous EPTS example, say that a coach wanted to evaluate the physical performance of a player in a match. A fundamental question relating to the content validity of the EPTS system would be the extent to which it could capture the entirety of the construct of 'physical performance'. To answer the question, the coach would first need to list all of the characteristics that they deem important in a player with respect to the physical component of competition. Whilst the EPTS system may provide suitable data by which to evaluate the velocity and potentially the acceleration performance of the player, it may be less suitable for understanding how the athlete jumped in heading contests, or the strength they exhibited whilst grappling with their opponent. Thus, in order to fully evaluate physical performance, other technologies or assessments would be required (such as inertial measurement units and visual analysis) and the system would not be deemed to show complete content validity for measuring the particular construct of 'physical performance'. Thus, content validity should not always be considered as a measure of the 'quality' of a technology per se, for some technologies are not designed to measure entire constructs. Of course, it is worth noting in this example that the validity of the construct itself also warrants investigation. For instance, such an investigation does not reveal the extent to which physical performance is actually important to overall 'football performance'. This is a separate question that is typically not tied directly to the technology itself. Further, what may be portrayed as incomplete content validity in one sport may be all that is required in another; for instance, a power meter attached to a bicycle may provide all the relevant information to a coach with respect to physical output in that given sport, but would obviously be inappropriate for use in most team sports.

- *Predictive:* Prediction is often considered the 'Holy Grail' of many scientific pursuits, and in sports technology it is no different. As the name implies, this form of validity refers to the ability of the technology to be used for predictive purposes (Streiner, et al., 2015). Multiple examples exist within sports domains across a variety of

applications. For instance, wearable sensors utilise auditory or haptic feedback on an athlete's technique which is based on them displaying a specific movement pattern that is predicted to correspond to a 'good' outcome (Düking, et al., 2017). Examples are as broad as sports data models being used to predict competition outcomes in almost any sport (Haghighat, et al., 2013) to the prediction of friction forces in footwear based on soil composition in natural turf (Guisasola, et al., 2009). The increased widespread utility of machine learning methods has seen a concomitant increase in exploration of technology application for predictive purposes. However, this has often been accompanied by a lack of understanding with respect to how such models work (Robertson, et al., 2017). Thus, although predictive validity can still be considered as one of the most valuable measures of quality for a technology, it is also important that the end user has some understanding with respect to how the prediction originated.

Other types of validity such as discriminant and convergent also exist; however, due to their comparatively lower relevance to technology evaluation, these are not discussed here. Readers are referred to (Streiner, et al., 2015) for further reading on the topic.

2.1.3 Reliability

Re-test reliability is perhaps the most commonly undertaken assessment of quality for technologies in sport. It refers to the consistency of output produced by a technology over multiple recorded instances (Terwee, et al., 2010). These instances are typically spaced over relatively short time periods; typically, minutes or hours, or sometimes days. Re-test reliability considers differences between test sessions that originate both systematically (as a function of the technology) and randomly (as a function of the environment and/or user). Thus, although it serves to provide a good overall representation of variability in the tested technology, it does not necessarily identify the source of where this variability lies.

Repeatability and reproducibility are two terms that are often used to separate the systematic and random contributors to any between-test differences (Bartlett & Frost, 2008). The former refers to what many would consider the traditional, laboratory-based understanding of reliability: the extent to which the output of the technology is repeatable given the exact same conditions upon re-test. Consequently, repeatability is of particular importance in highly repetitive scenarios conducted in the same environment. The latter refers to the extent to which the technology is reliable once tested across a range of individuals, scenarios and sports. Although establishing repeatability may be an essential

first step for many technologies, for those seeking sustained and widespread application into the broader sporting industry, reproducibility is absolutely crucial.

Inter-rater and intra-rater reliability are two further forms that are particularly relevant to technologies that require human involvement in their operation. Given that both forms relate to the generation of random differences in repeated testing of a technology, these properties show some relationship with reproducibility. The former, inter-rater reliability, refers to the level of agreement in output of the technology following repeated use by two or more different individuals (Baumgartner & Jackson, 1998). Thus, it is of particular relevance to technologies that are likely to experience use by different individuals within an organisation and in such situations, whereas the quality of the use could differ as a function of an individuals' level of training or experience. The latter, intra-rater reliability, can essentially be considered a sub-component of reproducibility, in that it relates to the variability in output following repeated use of a technology by the same person (Baumgartner & Jackson, 1998). Although this could be considered a representation of the level of skill and consistency of the individual themselves, more complex technology may be more difficult to use consistently – even when it is utilised by someone well trained in its operation.

A final consideration relating to reliability of technology is stability. Stability is similar to re-test reliability, though there is a key distinction which relates to the longer time period, i.e., its re-test reliability over many repeated assessments, over months and years (Schmidt, et al., 2003). Stability is of particular relevance to technology that may be expected to change or degrade over time after repeated use, which is the case of machinery or equipment requiring calibration.

Some examples of sport technology evaluations of reliability in the literature can be found below:

- The reliability of five different technologies used to measure bar velocity in weightlifting (Courel-Ibáñez, et al., 2019)
- Inter and intra-device reliability of an accelerometer used in player tracking in team sports (Nicolella, et al., 2018)
- Inter and intra-rater reliability of live football statistics captured from a commercial operator (Liu, et al., 2013)

2.1.4 Responsiveness & Sensitivity

Responsiveness of a technology can be defined as its ability to detect worthwhile and 'real' change (Beaton, et al., 2001). Thus, this property is less relevant for technologies delivering raw output to a user and more important for those that undertake some level of analysis or interpretation of the data prior to delivering the user. As it considers the actual application of the technology in terms of its analysis and interpretation, it also serves to provide an evaluation of the technology-human interaction. Considerable debate exists with respect to the differences (or non-differences) between terms such as responsiveness, sensitivity and minimum important difference. Thus, splitting responsiveness into three main facets is recommended when considered with respect to the quality of a technology. Such facets are who which and what (Beaton, et al., 2001).

- *Who:* This relates to whether the output of the technology considered over time relates to an individual or a group. This is an important distinction from an interpretation perspective; humans tend to undervalue changes observed at the individual level as opposed to those of a group. For instance, a 5-mm decrease in diastolic blood pressure due to the influence of a given technology may not be considered as meaningful for a single individual; on the other hand, when observed as a mean change across a group of individuals it may well be considered (Deyo & Patrick, 1995).
- *Which:* This refers to the timing of any comparisons made between outputs from a technology. Ideally, changes in an output over time would be observed in the same individual or group of individuals. However, this is not always feasible, particularly with repeated data collections over longer time periods (i.e., in professional sports, rosters or squads of players may turn over substantially from year to year). Sometimes, hybrid approaches that combine comparisons both across and between the same individuals may be used.
- *What:* This refers to the nature of the change that can be detected by the technology. Thus, facet in and of itself can be split into five types. There exist various equations that can be used to quantitatively define each of the five; however, in the interest of brevity these are not provided here (see Beaton, et al., 2001 for further reading).
    - *The minimum detectable change capabilities of the technology:* This essentially considers the resolution of the data output. For instance, a timing device that can measure to an accuracy of 1/10 of a second may be sufficiently appropriate for use in timing an athlete's lap time in training. However, it will not be feasible for determining the outcome of a 100 m final in a World Championships.
    - *The minimum actual change capabilities of the technology:* This is almost the same as the previous type, but it also includes a consideration of the error

inherent in the technology (when known.) It is sometimes referred to as the 'minimally detectable change'.

- *The change observed between measurements using the technology in a given population:* It includes both positive and negative changes resulting as a consequence of the technology and considers the change across the group rather than individuals. It is sometimes referred to as 'observed change.'
- *Observed change in a population that has improved:* This defines the characteristics of a given population that improved as a consequence of the technology. For instance, this could refer to a scenario whereby a group of injured athletes using a certain technology returned to baseline muscle strength measures more rapidly than normal within a given time period. It is sometimes referred to as 'estimated change.'
- *Observed change in a population that has meaningfully improved:* As per the scenario above, it could also refer to only assessing those athletes who not only returned to baseline muscle strength, but also safely returned to training or competition. It is sometimes referred to as 'important change.'

This module has summarised the key properties pertaining to the quality of a technology as it relates to application in sport. Even though many of the items will be relevant to any technology, it is worth noting that some are more specific to certain types, contexts or applications than others. An operational framework whereby these indicators of quality can be utilised practically in the field is presented in Theme 4.1.1.

2.2 Feasibility

Feasibility characteristics of a technology relate to the ease and convenience in which the technology can be utilised in the practical environment. Such characteristics of quality are often considered secondary to those discussed above, those which emanate from 'harder' science. However, with a growing proportion of technology moving from confinement to the laboratory to use the field, combined with the increased competition in the

marketplace, feasibility of technology has rapidly become an important consideration for sporting organisations. Below, four main considerations are discussed.

2.2.1 Complexity

In order to experience successful and ongoing implementation in the field, ideally the complexity of technology should be minimised. This can include a variety of considerations, depending on the nature and regularity of use of the technology:

- *Set-up and maintenance:* From a sports organisational perspective, often multiple training and competition locations will exist. In such cases, a simple set-up and maintenance procedure helps to ensure consistency in the output between multiple versions of the same technology. In situations where a single version of the technology exists, reducing complexity becomes even more important, as the system or device may need to be packed, transported, unpacked and installed simply for rapid use. Reduced cabling, wireless communication and increased battery life represent some examples of how this process can be made less complex.
- *Extraction and utility of data and output:* Many technologies are accompanied by software incorporating graphical user interfaces (GUI) in order to facilitate user-friendly examination of output. Some of these GUIs will be customisable at the user end. However, extraction of the data in a raw format is not always available. This is problematic, since the user may require this for a variety of reasons, such as to conduct their own evaluation into the quality of the data, or to apply it to their own analysis, or to repurpose the data for multiple uses (see 'Flexibility' below.)
- *Integration:* This refers to the ability for a technology to be integrated with either existing or new processes within the organisation. It is often a difficult area for technology manufacturers to emphasise in their development, as it requires foresight with respect to upcoming trends in the sport industry and also ongoing development to facilitate such integrations. Examples could include combining kinetic data from a force plate with 3D kinematic data obtained from another manufacturer, or combining match event and EPTS data in team sports.
- *Flexibility:* This has become an increasingly important area for manufacturers and is discussed in greater detail in Theme 2.2.4 below. Above, the role of GUIs in providing user customisation and flexible visualisation options was discussed. However, this consideration may sometimes extend to all components of a technology. For example, some technologies utilised in live-sports competition may require near real-time output available in different formats depending on whether the content is required by a coach, the media, broadcaster or adjudicator. Sometimes multiple

formats of these data may be required, using different filters or processing techniques.

- *Multicollinearity:* It is a term typically used in statistics, which refers to two (or more) variables in a model that are strongly associated with one another (Alin, 2010). Often a resulting action from such scenarios is that one or more of the variables is discarded, in order to create a more simple solution using a single variable. This is important in technologies that output multiple variables as part of their basic function, which can often cause issues for the user. For example, a force plate manufacturer that outputs 10 variables in their software which all report on similar constructs (i.e., derivatives of force) may cause confusion in the user as to how many should be used to inform processes and decisions. Including more variables may also result in the user overemphasising the output from the device altogether, potentially at the expense of other information.

Of course, not all technology can always be easy to implement; sometimes, in early stages of new innovation, it can be anything but easy. However, manufacturers emphasising the above characteristics in their product development will tend to experience success in the form of positive client feedback and ongoing usage.

2.2.2 Cost

Cost is a self-explanatory concept, but with respect to evaluating technology it refers to more than just money. Cost also relates to the time investment required both prior to utilising a technology and what is required on an ongoing capacity. For the former, this can involve the time, financial outlay and relevant expertise in assessing the market for the most suitable selection. If the market of a particular type of technology is especially saturated, this has the potential to be an overwhelming and elongated process. Besides, if the technology is complex or new, it might require the expertise and involvement of the most experienced staff within an organisation. For the latter, some technologies may require ongoing servicing and quality checks which can also incur cost. In technologies producing large amounts of data, it may also correspond to ongoing cost around storing and processing data through computational means.

Often an assessment of cost is undertaken against either the overall benefit of the organisation or the expectations of key stakeholders. For the former, this may often be an exercise that involves multiple stakeholders. Such is the case in team sports, where the technology may be shared across men and women teams, or youth and senior teams, one of which may find greater value than the others, thus changing the assessment of cost. For

example, an organisation may be more willing to spend on a technology to be used with a senior team than a youth one. A similar scenario in institute or university sport programs could take place when the benefit is higher to one ability level or sport over another. If the new technology is replacing multiple other technologies, then there may be less resistance to a higher initial cost outlay. Moreover, when the technology is being utilised as a supplementary information source to an existing judgement or process, its impact can sometimes be simply confirmatory. Therefore, given these considerations and the involvement of stakeholder time and expertise in cost assessment, it can be difficult to precisely determine its evaluation in quantitative terms.

From an expectation perspective, often internal 'public relations' exercise may be required within organisations, in order to educate financial controllers on the benefits of a proposed investment or purchase. Again, technologies that display widespread use across an organisation, or facilitate competitive advantage, may be easier to gain support for. This is often also related to any preconceived determinations of what value the technology will provide. When the value is spread over a time frame or in the longer term, humans often struggle to value its impact as highly as when the influence is felt more in the acute, short term. Organisations that are well-structured and mature with respect to technology use and innovation are typically more likely to engage a technology, due to its strategic value rather than a will to receive a more tangible return of investment.

2.2.3 Training required

The level of training required to manage and operate various technologies in contemporary sport is a non-trivial issue. With new technologies forming part of the daily operations of many sporting organisations, many will require multiple staff to be upskilled in their installation, application and troubleshooting, in order to see their full value. As discussed in Theme 1.2.2, obtaining a balancing act for staff with respect to their time spent upskilling in new technology against their existing expertise is a difficult exercise. In some roles within sport (i.e., performance analysis or information technology), developing an increased aptitude in emerging technologies is often considered part of the role. However, for coaching and administration teams, sometimes a 'champion' individual is required to take the lead. This also extends to the governance of data emanating from technology, which, as discussed earlier in the course, organisations have an ethical responsibility to uphold.

Spending stakeholder time on training can of course be positive, particularly if the skills developed are also transferable to other tasks. For example, technology that delivers output via open source programming languages not only breeds a level of independence in users,

but it also serves as a form of crowdsourcing for manufacturers, essentially performing as a research and development service. Some technology manufacturers offer training as part of an ongoing service to users, and this can be seen as a value-add and sometimes as a market differentiator for organisations faced with a choice about which manufacturer to work with. A balancing act is again required by manufacturers in order to ensure that users can rapidly and easily upskill in any new and essential changes to aspects of their technology, but not make so many alterations that it becomes difficult for key stakeholders to remain up-to-date. In summary, technologies facilitate easy use with minimal training initially, but still provide the user with multiple and flexible courses of action, as they develop their expertise is typically well-rated by this criterion.

2.2.4 Interpretability

This feature relates to the ability to which practical meaning can be attributed to output from a technology (Lohr, 2002). It bears some relationship with complexity, given that it includes consideration of a technology output to be visualised, understood and acted upon appropriately. The flexibility of GUIs was discussed above but it is also important to recognise that technologies incorporating visualisations or reports via software should be delivered in a default form that is designed using well-established psychophysics principles. Psychophysics refers to how sensory mechanisms interact with perceptual processes (Stevens, 2017) and, thus, it is particularly relevant with respect to how output of a technology is perceived and interpreted by a human user. As sport environments are generally fast paced, translating complex information through visualisations instead of via pages of data or text reports is recommended where possible, as this helps to offload cognitive work by the user to automatic perceptual processing (Kale, et al., 2018). This has the intended benefit of facilitating such interpretations to be made more rapidly and, therefore, also actioned faster (Larkin & Simon, 1987). Where relevant, any information output should communicate any uncertainty or error as part of its reporting (Kay, et al., 2016). Other considerations relating to visual presentation of data from technology include informational and computational equivalence. The former relates to how a given visualisation can impart the same level of information as those included in a written report, or data table (Larkin & Simon, 1987). The type, size, and colour of a visualisation are all important features that can improve this characteristic. The latter relates to the ability for more detailed data to be processed, analysed and reported on rapidly at the same (or acceptably similar) rate, as more basic metadata (Larkin & Simon, 1987).

Another important part of interpretability is ensuring that the actual interpretations which the technology facilitates are correct. Outputs or reports from a technology that cause for

incorrect conclusions to be drawn by the user, whether intentional or not, can be a major problem in the field when stakeholders are not equipped with the relevant skills to question the origin of the information. Put simply, interpretability precedes precision. This means that in many sport scenarios the extent for key stakeholders to understand data or an analytical model is arguably even more important than its accuracy (provided, of course, the accuracy is not completely poor). Moreover, both accuracy and interpretability exist on a continuum and are not dichotomous. But, if given the choice between a slightly less accurate technology that is actionable by users in the field and one that is accurate but not implementable, the decision is easy.

Similar to cost, quantitatively rating interpretability can be a difficult exercise. Typically, the process of evaluating a technology's interpretability might be done subjectively by staff, either in the form of Likert-type ratings or through subjective means.

References

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(3), 370-374. https://doi.org/10.1002/wics.84

Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound in Obstetrics and Gynecology: *The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology, 31*(4), 466-475. **https://doi.org/10.1002/uog.5256**

Baumgartner, T. A., & Jackson, A. S. (1998). *Measurement for evaluation in physical education and exercise science* (6th ed.). WCB/McGraw-Hill.

Beaton, D. E., Bombardier, C., Katz, J. N., & Wright, J. G. (2001). A taxonomy for responsiveness. *Journal of clinical epidemiology, 54*(12), 1204-1217. https://doi.org/10.1016/S0895-4356(01)00407-3

Courel-Ibáñez, J., Martínez-Cava, A., Morán-Navarro, R., Escribano-Peñas, P., Chavarren-Cabrero, J., González-Badillo, J. J., & Pallarés, J. G. (2019). Reproducibility and repeatability of five different technologies for bar velocity measurement in resistance training. *Annals of biomedical engineering, 47*(7), 1523-1538. https://doi.org/10.1007/s10439-019-02265-6

Currell, K., & Jeukendrup, A. E. (2008). Validity, reliability and sensitivity of measures of sporting performance. *Sports medicine, 38*(4), 297-316. https://doi.org/10.2165/00007256-200838040-00003

Deyo, R. A., & Patrick, D. L. (1995). The significance of treatment effects: the clinical perspective. *Medical care, 33*(4), AS286-AS291.

Düking, P., Holmberg, H. C., & Sperlich, B. (2017). Instant biofeedback provided by wearable sensor technology can help to optimize exercise and prevent injury and overuse. *Frontiers in physiology, 8*, 167. https://doi.org/10.3389/fphys.2017.00167

Guisasola, I., James, I., Llewellyn, C., Stiles, V., & Dixon, S. (2009). Quasi-static mechanical behaviour of soils used for natural turf sports surfaces and stud force prediction. *Sports Engineering, 12*(2), 99-109. https://doi.org/10.1007/s12283-010-0035-2

Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal, 2*(5), 7-12. https://www.researchgate.net/publication/262560138_A_Review_of_Data_Mining_Techniques_for_Result_Prediction_in_Sports

Kale, A., Nguyen, F., Kay, M., & Hullman, J. (2018). Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE transactions on*

visualization and computer graphics, 25(1), 892-902. https://doi.org/10.1109/TVCG.2018.2864909

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. *ACM,* (pp. 5092-5103). https://doi.org/10.1145/2858036.2858558

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive science, 11*(1), 65-100. https://doi.org/10.1111/j.1551-6708.1987.tb00863.x

Lohr, K. N. (2002). Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of Life Research, 11*(3), 193-205. https://doi.org/10.1023/a:1015291021312

Liu, H., Hopkins, W., Gómez, A. M., & Molinuevo, S. J. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. *International Journal of Performance Analysis in Sport, 13*(3), 803-821. https://doi.org/10.1080/24748668.2013.11868690

Nicolella, D. P., Torres-Ronda, L., Saylor, K. J., & Schelling, X. (2018). Validity and reliability of an accelerometer-based player tracking device. *PloS one, 13*(2), e0191823. https://doi.org/10.1371/journal.pone.0191823

Robertson, S. J., Burnett, A. F., & Cochrane, J. (2014). Tests examining skill outcomes in sport: a systematic review of measurement properties and feasibility. *Sports Medicine, 44(*4), 501-518. https://doi.org/10.1007/s40279-013-0131-0 Robertson, S., Bartlett, J. D., & Gastin, P. B. (2017). Red, amber, or green? Athlete monitoring in team sport: the need for decision-support systems. *International journal of sports physiology and performance, 12*(s2), S2-73. https://doi.org/10.1123/ijspp.2016-0541

Robertson, S., Kremer, P., Aisbett, B., Tran, J., & Cerin, E. (2017). Consensus on measurement properties and feasibility of performance tests for the exercise and sport sciences: a Delphi study. *Sports Medicine-Open, 3(*2), 1-10. https://dx.doi.org/10.1186%2Fs40798-016-0071-y

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods, 8*(2), 206. https://psycnet.apa.org/doi/10.1037/1082-989X.8.2.206

Stevens, S. S. (2017). *Psychophysics: Introduction to its perceptual, neural and social prospects*. Routledge.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use.* Oxford University Press.

Terwee, C. B., Mokkink, L. B., van Poppel, M. N., Chinapaw, M. J., van Mechelen, W., & de Vet, H. C. (2010). Qualitative attributes and measurement properties of physical activity questionnaires. *Sports Medicine, 40*(7), 525-537. https://doi.org/10.2165/11531370-000000000-00000