Sample size calculations for educational interventions: principles and methods

Meghan M. McConnell, PhD · Sandra Monteiro, PhD · Gregory L. Bryson, MD, FRCPC, MSc

Received: 24 January 2019 / Revised: 15 April 2019 / Accepted: 15 April 2019 / Published online: 22 May 2019
© Canadian Anesthesiologists' Society 2019

Scenario

You are your department's director of obstetrical anesthesia and you plan to introduce programmed intermittent epidural bolus analgesia to your hospital. In doing so, you will need to teach your anesthetic and nursing colleagues about the rationale for this technique as well as how to program and use the new required pumps. You plan a mixture of teaching rounds and online learning modules. Building on research showing that repeated testing can enhance long-term retention of information,¹ you plan to integrate weekly quizzes consisting of different question formats (e.g., short-answers, multiple-choice

items). To determine whether repeated quizzes promote learning, you would like to conduct an experiment, where half of the individuals receive weekly quizzes and the other half receive additional study material. One month after the completion of the educational program, you would like to test your colleagues' knowledge of the new technique to determine whether weekly quizzes enhanced retention of information presented. You must now determine how many individuals you will need to be sure your intervention has worked.

Introduction

Successful healthcare systems depend on efficient education of healthcare professionals—not only for undergraduate, resident, and fellow trainees but also for the continuing professional development of clinicians in practice. The Best Evidence in Medical Education movement emphasizes the importance of providing clinical educators with an evidence base for medical education scholarship and practice to answer questions related to what works, in what context, with which groups, and at what cost.² As with any research endeavour, it is important to determine the sample size required to find a difference between the experimental and control groups. While there are many articles describing sample size calculations within clinical research,³ the purpose of this editorial is to discuss sample size calculations within the context of medical education interventions. We specifically focus on sample size calculations for comparing different groups in which the outcome is a continuous (interval or ratio) dependent variable, as such interventional designs are common practice in medical education research.

M. M. McConnell, PhD (✉)
Department of Innovation in Medical Education, University of
Ottawa, Ottawa, ON, Canada
e-mail: meghan.mcconnell@uottawa.ca

Department of Anesthesiology and Pain Medicine, University of
Ottawa, Ottawa, ON, Canada

Faculty of Medicine, University of Ottawa, 850 Peter Morand
Crescent – 102, Ottawa, ON K1G 5Z3, Canada

S. Monteiro, PhD
Department of Health Research Methods, Evidence and Impact,
McMaster University, Hamilton, ON, Canada

Department of Anesthesiology, McMaster University, Hamilton,
ON, Canada

G. L. Bryson, MD, FRCPC, MSc
Department of Anesthesiology and Pain Medicine, University of
Ottawa, Ottawa, ON, Canada

Clinical Epidemiology Program, Ottawa Hospital Research
Institute, Ottawa, ON, Canada

Parameters for sample size calculations

The primary goal of sample size calculations is to determine the number of participants required to measure the effects of an intervention on a particular outcome or an association between variables.⁴ Sample size is a primary consideration in the planning stage of any study and informs many aspects of trial design and conduct. Determining the appropriate sample size typically depends on four parameters: significance criterion, desired statistical power, anticipated difference in scores, and estimated measurement variability.

1. The *significance criterion* is the probability of committing a type-I error, that is, the likelihood of incorrectly finding a statistically significant effect when, in reality, no such effect exists. This probability is referred to as alpha (α) and serves as the comparison point for the *P* values of statistical tests when determining the significance of the relevant findings (i.e., whether or not to reject the null hypothesis). Conventionally, the significance criterion is set at 5% ($\alpha = 0.05$). As this value decreases, the sample size needed to detect a significant effect increases.
2. *Statistical power* is the probability of finding a statistically significant effect when such an effect is present. In this way, statistical power is analogous to the sensitivity of a diagnostic test³ and is the complement of the probability type-II error. Power is conventionally set at a minimum of 80% (0.80), meaning that 80% of the time, a researcher will find a true statistically significant effect. While high power is always desired from a study, as power increases, so does the required sample size.
3. The *anticipated difference*, often referred to as the minimally important difference (MID), is the smallest measured difference between comparison groups that the researcher would like to detect. Within clinical research, the anticipated difference represents “the smallest difference in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s management”.⁵ The MID may be expressed as a difference in means, ranks, or a desired correlation depending upon the measure of the outcome of interest. Within education research, the anticipated difference is the smallest difference in the outcome measure (e.g., test score, time taken to complete procedure, etc.) that educators and/or trainees would perceive as beneficial and would warrant potential changes in curriculum design. As the anticipated

difference decreases, the sample size required to detect a statistical difference increases.

4. And finally, sample size calculations also require an *estimate of the variability* (e.g., standard deviation) of the outcome measure. As variability increases, the sample size required to detect the MID increases. Estimates of measurement variability are ideally determined by collecting preliminary data using similar study populations, or through reviewing the literature.

With knowledge of these parameters, researchers can calculate sample size requirements. The mathematical formulae for calculating sample size can be found in the Appendix. While statistical significance and power are typically set by convention (e.g., ≤ 0.05 and ≥ 0.80 , respectively), researchers typically need to estimate the anticipated difference of the effect (e.g., educational intervention) and the variability of the measured outcome (e.g., test score). Such estimates may come from a variety of sources, including previously published data from pilot studies, conference presentations, unpublished manuscripts, and so forth.

Sample size calculations in education research

While precise mathematical formulae are available for calculating sample sizes (Appendix), these depend on specifying values for inputs, such as the anticipated treatment effect and standard deviation. In practice, there is rarely sufficient information to accurately estimate such values in medical education research. One reason for this is related to the types of outcome measures used in education research. Educational interventions often focus on impacting latent constructs, which are theoretical in nature and cannot be observed or measured directly⁶; for example, patient satisfaction, resident communication, or physician well-being. Identifying outcome measures associated with such latent constructs is arguably more difficult than identifying physical measures (e.g., average systolic blood pressure values). Moreover, variations in scales used and types of data collected present challenges when attempting to draw comparisons. For example, a variety of validated communication scales exist in the literature, making it quite challenging to estimate the anticipated difference (e.g., a ten-point difference on a communication rating scale between the education intervention vs control group) or the variability of the outcome measure (e.g., ± 2 on a Likert rating scale).

For this reason, education researchers often promote the use of effect sizes in calculating sample size requirements. Standardized effect sizes provide an estimate of the

Figure Relationship between sample size and effect size.
Relation entre la taille d'échantillon et la taille d'effet.

magnitude of an effect or association measured in statistical units (e.g., standard deviations or percentages of total variance), meaning they provide standardized numbers that are not tied to a specific measurement tool⁴ (see Appendix). Generally speaking, the larger the effect size, the smaller the sample size requirements (Figure). And while various estimates of effect sizes exist depending on study design (e.g., Cohen's *d*, Glass's Δ , Hedges' *g*), their standardized nature has led to the classification of effect sizes of 0.20, 0.50, and 0.80, as "small", "medium", and "large", respectively.⁷

Effect sizes are particularly useful in education research contexts, as their standardized nature allows researchers to quantitatively compare results from different studies using disparate scales and measures.⁸ Consequently, reporting effect sizes and their corresponding confidence intervals is becoming common practice in educational literature, as it allows "for a more communal knowledge base....so that the results from different studies can be compared more readily".⁹ Given the relevance of effect sizes in education research, various studies have documented effect sizes for educational and psychological interventions. For example, Hattie¹⁰ recently conducted a synthesis of 1,200 meta-analyses and reported mean effect sizes for 195 different educational activities, such as classroom discussion (mean effect size, $d = 0.82$), feedback ($d = 0.73$), formative evaluation ($d = 0.68$), self-questioning ($d = 0.64$), small group learning ($d = 0.47$), and problem based learning ($d = 0.12$) to name a few.

Based on published effect sizes and historical norms obtained from various educational studies, several researchers have suggested sample size guidelines for education research. For example, Cohen¹¹ proposed that for small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effect sizes, the norm for sample sizes would be 400, 64, and 25, respectively (see Appendix 1 for calculations). There are also a variety of online resources and computer

programs that have been developed for calculating appropriate sample sizes.^{12,13}

Mishaps to avoid when using effect size for sample size calculations

By this point, we have highlighted the usefulness of effect size estimates in calculating sample size requirements in education research. That being said, we also discuss some of the limitations to this approach.

First and foremost, researchers should avoid estimating sample size using designations of small, medium, or large effect sizes (e.g., find the sample size needed to detect a "medium" effect); after all, these designations fail to take into account other variables such as the accuracy of the assessment instrument and the characteristics of the study population.⁸ Instead, when using effect size estimates in sample size calculations, researchers should use norms within research communities to identify representative and expected normative values of effect size.¹⁴

Second, as discussed above, one of the reasons educational researchers emphasize the usefulness of effect size in sample size calculations is related to the types of outcome measures used. Educational research typically involves dependent measures that cannot be easily observed, such as clinical or procedural knowledge, empathy, communication skills, and so on. Many of these educational outcomes are often developed locally by the research team and consequently, vary from institute to institute. It is for this reason that prominent medical education researchers highly recommend that researchers avoid developing new, institute-specific measurement tools unless it is absolutely necessary. Just as clinical research requires careful consideration of the most appropriate outcome measures based on research question(s) and study design, the same care must be taken when identifying the appropriate outcomes for the educational constructs of interest. Whenever possible, use instruments that have been previously validated in the literature. In instances where researchers feel the need to develop a new tool, such tools must not only be constructed and calibrated in a specific way, but reliability and validity evidence must also be collected prior to using such a tool in an interventional study.⁶ Any researcher applying an inappropriate or inaccurate tool will be challenged to find an effect, regardless of the estimated effect size.

Third, as with clinical research, it is important to consider potential attrition rates when determining the necessary sample size for educational studies. Such dropout issues are typically more important in longitudinal study designs, and thus, it is important to plan sample size carefully in such cases. Fourth, it is important to avoid conflating the magnitude of an effect

size with the actual importance of the effect. While educational interventions typically have larger effect sizes than clinical interventions, this is not to say that the former is more relevant. To paraphrase Wilkinson, small replicable effects involving meaningful outcomes can be particularly important, while large effects involving trivial outcomes may be less important.¹⁵

Lastly, effect size estimates are often associated with some level of uncertainty coming from numerous factors, such as the educational background of the students, random variation in mood and attention, and other factors. As a result, there is consensus within the educational and social sciences community to report confidence intervals associated with effect sizes whenever possible.⁹ Confidence intervals provide a range of possible values for the parameter of interest (in this case, effect size) with some degree of confidence. Providing confidence intervals for effect sizes emphasizes not only the magnitude of the effect, but also provides information regarding the precision of the estimate. Imprecision around the effect size estimate can have a deleterious effect on sample size calculations.

Conclusions

As with clinical research, sample size is an important part of designing and performing high quality education research. Within healthcare education settings, the use of effect size in sample size calculations is well established. Considering sample size early on in the study phase allows for more meaningful data that will ultimately have a higher impact on educational practice. In some instances, multicentre studies will be necessary, but it is much better to determine this early on in the study design process as it will save much time and frustration later on.

Resolution of the scenario

Referring back to the scenario presented at the beginning of this article, you are interested in determining how many individuals would be needed to determine whether weekly mixed-format quizzes (e.g., short-answer, multiple-choice items) promote long-term retention of knowledge of intermittent bolus analgesia. After reviewing the literature, you were not able to find established measures of variability and anticipated differences between the two groups (e.g., repeated quizzes vs repeated study). Nevertheless, a recent meta-analysis of educational interventions in higher education revealed that mixed test formats produced large effect sizes (Hedges's $g = 0.81$).¹⁶ Using the shortened version of the formula¹⁷ illustrated in the Appendix:

$$n = 16 \frac{1}{\text{effect size}^2} = 16 \frac{1}{0.81^2} = 16 \frac{1}{0.66} = 24.24$$

Solving the formula gives $n = 24.24$ or 25 in each group, resulting in a total sample size of 50 learners.

Le calcul des tailles d'échantillons pour les interventions éducationnelles: principes et méthodes

Mise en situation

Vous êtes le directeur d'anesthésie obstétricale de votre département et vous avez l'intention d'introduire l'analgésie par administration programmée de bolus épiduraux dans votre institution. Pour ce faire, il vous faudra enseigner à vos collègues en anesthésie et en soins infirmiers les fondements de cette technique ainsi que la façon de programmer et d'utiliser les nouvelles pompes nécessaires à ce type d'analgésie. Vous planifiez d'organiser une combinaison de présentations d'enseignement et de modules d'apprentissage en ligne. En vous appuyant sur des recherches démontrant que des tests répétés peuvent améliorer la rétention d'informations à long terme,¹ vous avez l'intention d'incorporer des examens hebdomadaires comportant différents formats de question (par exemple, réponses courtes, questionnaires à choix multiples). Afin de déterminer si les examens répétés promeuvent l'apprentissage, vous aimeriez réaliser une expérience dans laquelle la moitié des participants passent des examens hebdomadaires et l'autre moitié reçoit du matériel d'étude supplémentaire. Un mois après la fin du programme de formation, vous aimeriez tester les connaissances de vos collègues concernant la nouvelle technique afin de déterminer si les examens hebdomadaires ont amélioré la rétention des informations présentées. Vous devez désormais déterminer le nombre de personnes dont vous aurez besoin afin de vous assurer que votre intervention a fonctionné.

Introduction

Les systèmes de soins de santé performants dépendent d'une formation rigoureuse des professionnels de la santé – pas seulement des étudiants, résidents et stagiaires, mais

également en matière de développement professionnel continu des cliniciens déjà en pratique. Le mouvement des Meilleures données probantes en formation médicale (*Best Evidence in Medical Education*) souligne l'importance de procurer aux enseignants cliniques une base de données probantes favorisant l'érudition en formation médicale et dans la pratique afin de répondre aux questions telles que: Qu'est-ce qui fonctionne? Dans quel contexte? Avec quels groupes? Et à quel coût? ² Tout comme c'est le cas dans n'importe quel effort de recherche, il est essentiel de déterminer la taille d'échantillon nécessaire à déceler une différence entre les groupes intervention et témoin. Bien qu'il existe de nombreux articles portant sur le calcul de la taille d'échantillon en recherche clinique, ³ l'objectif de cet éditorial est de discuter le calcul de la taille d'échantillon dans un contexte d'interventions de formation médicale. Nous mettons spécifiquement l'emphasis sur le calcul de la taille d'échantillon lors de la comparaison entre différents groupes dans lequel le résultat est une variable dépendante continue (intervalle ou ratio), étant donné que des tels modèles d'intervention sont pratique courante en recherche sur la formation médicale.

Paramètres pour le calcul de la taille d'échantillon

L'objectif principal du calcul de la taille d'échantillon est de déterminer le nombre de participants requis pour mesurer les effets d'une intervention sur un résultat en particulier ou une association entre les variables. ⁴ La taille d'échantillon est un élément majeur de la phase de planification de toute étude et a un impact sur de nombreux aspects tant de conception que de réalisation d'une étude. La détermination d'une taille d'échantillon adaptée dépend généralement de quatre paramètres : le niveau de confiance (ou significatif), la puissance statistique souhaitée, la différence anticipée dans les scores, et la variabilité de la mesure estimée.

1. Le *niveau de confiance* est la probabilité de commettre une erreur de type I, soit la probabilité de trouver de manière erronée un effet statistiquement important, alors qu'en réalité un tel effet n'existe pas. On parle d'alpha (α) lorsqu'on fait référence à cette probabilité, et elle sert de point de comparaison pour les valeurs *P* des tests statistiques lorsqu'on détermine l'importance des résultats pertinents (c.-à-d. s'il faut rejeter ou non l'hypothèse nulle). Traditionnellement, le niveau de confiance est établi à 5 % ($\alpha = 0,05$). Plus cette valeur baisse, plus la taille d'échantillon nécessaire pour déceler un effet significatif augmente.
2. La *puissance statistique* décrit la probabilité de déceler un effet statistiquement significatif lorsqu'un tel effet

est présent. De cette manière, la puissance statistique est semblable à la sensibilité d'un test diagnostique ³ et est complémentaire à l'erreur de probabilité de type II. La puissance est traditionnellement établie à un minimum de 80 % (0,80), ce qui signifie que 80 % du temps, un chercheur décelera un véritable effet statistiquement significatif. Étant donné qu'une puissance élevée est toujours souhaitable dans une étude, plus la puissance augmente, plus la taille d'échantillon nécessaire augmente.

3. La *différence anticipée*, qu'on nomme souvent différence minimale significative (DMS), est la plus petite différence mesurée entre les groupes que le chercheur aimerait déceler. En recherche clinique, la différence anticipée représente « la plus petite différence dans le domaine d'intérêt perçu par les patients comme avantageux et qui pourrait justifier, en l'absence d'effets secondaires significatifs et de coût excessif, un changement dans la prise en charge du patient ». ⁵ La DMS peut s'exprimer en tant que différences de moyennes, de rangs, ou comme une corrélation souhaitée selon la mesure du résultat auquel on s'intéresse. En recherche sur la formation, la différence anticipée représente la plus petite différence en ce qui touche au critère d'évaluation (par ex., notes à un examen, temps nécessaire pour réaliser l'intervention, etc.) qui serait perçue par les formateurs et/ou les stagiaires comme avantageuse et justifierait des changements potentiels à la conception du programme d'études. Plus la différence anticipée se réduit, plus la taille d'échantillon nécessaire à déceler une telle différence statistique augmente.
4. Enfin, le calcul de la taille d'échantillon nécessite également une *estimation de la variabilité* (par ex. écart type) du critère d'évaluation. Plus la variabilité augmente, plus la taille d'échantillon nécessaire à déceler la DMS augmente. Les valeurs estimées de la variabilité de mesure sont, dans l'idéal, mesurées en colligeant des données préliminaires portant sur des populations à l'étude semblables, ou par une revue de la littérature.

En connaissant ces paramètres, les chercheurs peuvent alors calculer les tailles d'échantillon nécessaires. Les formules mathématiques pour calculer les tailles d'échantillon sont présentées dans l'annexe. Alors que le niveau de confiance et la puissance sont en règle générale établis par convention (par ex., $\leq 0,05$ et $\geq 0,80$, respectivement), les chercheurs ont habituellement besoin d'estimer la différence anticipée de l'effet (par ex., d'une intervention éducationnelle) et la variabilité du critère à l'étude (par ex., la note à l'examen). De telles estimations peuvent provenir de diverses sources, notamment de

données publiées d'études pilote, de présentations lors de conférences, de manuscrits non publiés, etc.

Calculs des tailles d'échantillon en recherche sur l'éducation

Bien qu'il existe des formules mathématiques précises pour calculer les tailles d'échantillon (voir annexe), celles-ci dépendent de valeurs spécifiant les données saisies, comme par exemple l'effet anticipé d'un traitement et l'écart type. Dans la pratique, l'information disponible est rarement suffisante pour estimer précisément de telles valeurs dans la recherche en formation médicale. L'une des raisons expliquant cet écueil est lié au type de critères d'évaluation utilisés dans la recherche en éducation. Les interventions éducationnelles se concentrent souvent sur leur impact sur des concepts latents, lesquels sont par nature théoriques et ne peuvent être observés ou mesurés directement⁶; citons par exemple la satisfaction des patients, la communication avec les résidents, ou le bien-être des médecins. L'identification de critères d'évaluation associés à de tels concepts latents est indubitablement plus difficile que l'identification de mesures physiques (par ex. les valeurs de tension artérielle systolique moyennes). En outre, les variations en matière d'échelles et de types de données ajoutent encore aux défis lorsqu'on tente de faire une comparaison. Par exemple, il existe plusieurs échelles validées de communication dans la littérature, ce qui rend difficile l'estimation de la différence anticipée (par ex., une différence de dix points sur une échelle d'évaluation de la communication entre le groupe intervention et le groupe témoin) ou la variabilité du critère d'évaluation (par ex., ± 2 sur l'échelle d'évaluation de Likert).

Pour cette raison, les chercheurs en éducation préconisent souvent l'utilisation des tailles d'effet pour calculer les besoins en taille d'échantillon. Des tailles d'effet normalisées procurent une estimation de la magnitude d'un effet ou d'une association mesurée en unités statistiques (par ex. les écarts type ou les pourcentages de variance totale); elles offrent donc des nombres standardisés qui ne sont pas liés à un outil de mesure spécifique⁴ (voir annexe). En règle générale, plus la taille d'effet est importante, moins les besoins en taille d'échantillon sont grands (figure). En outre, bien qu'il existe plusieurs estimations de tailles d'effet selon la méthodologie retenue pour l'étude (par ex., le d de Cohen, le Δ de Glass, le g de Hedges), leur nature standardisée a entraîné la classification des tailles d'effets de 0,20, 0,50 et 0,80 en « petite », « moyenne », et « grande », respectivement.⁷

Les tailles d'effet sont tout particulièrement utiles dans les contextes de recherche en éducation, étant donné que

leur nature standardisée permet aux chercheurs de comparer, de manière quantitative, les résultats de différentes études utilisant des échelles et des mesures disparates.⁸ Par conséquent, la communication des tailles d'effet et des intervalles de confiance correspondants devient de plus en plus populaire dans le champ de la littérature éducationnelle, étant donné que cela « permet de créer une base de connaissance plus commune... de telle manière que les résultats de différentes études peuvent être plus facilement comparés ».⁹ Étant donné la pertinence des tailles d'effet dans la recherche en éducation, diverses études ont documenté les tailles d'effet destinées aux interventions éducationnelles et psychologiques. Par exemple, Hattie¹⁰ a récemment réalisé la synthèse de 1200 méta-analyses et rapporté les tailles d'effet moyennes de 195 différentes activités éducationnelles, telles que les discussions en classe (taille d'effet moyenne, $d = 0,82$), les rétroactions ($d = 0,73$), l'évaluation formative ($d = 0,68$), les questionnaires auto-administrés ($d = 0,64$), l'apprentissage en petits groupes ($d = 0,47$), et l'apprentissage par problèmes ($d = 0,12$), pour ne citer que ces exemples.

Sur la base des tailles d'effet publiées et des normes historiques obtenues à partir de diverses études éducationnelles, plusieurs chercheurs ont proposé des recommandations de tailles d'effets pour la recherche en éducation. Par exemple, Cohen¹¹ a proposé que, pour des tailles d'effet petite ($d = 0,2$), moyenne ($d = 0,5$) et grande ($d = 0,8$), la norme pour les tailles d'échantillon soit de 400, 64 et 25, respectivement (voir l'annexe 1 pour le calcul). Il existe également plusieurs ressources en ligne et programmes informatiques mis au point pour calculer des tailles d'échantillon adaptées.^{12,13}

Pièges à éviter lors de l'utilisation de la taille d'effet pour le calcul de la taille d'échantillon

Nous avons souligné l'utilité des estimations de taille d'effet pour calculer les besoins en taille d'échantillon en recherche sur l'éducation. Cela étant dit, nous devons également aborder certains des écueils liés à cette approche.

En premier lieu, les chercheurs devraient éviter d'estimer la taille d'échantillon en utilisant les désignations de tailles d'effet petite, moyenne ou grande (par ex., trouver la taille d'échantillon nécessaire à détecter un effet « moyen »); après tout, ces désignations ne tiennent pas compte d'autres variables telles que la précision de l'instrument d'évaluation et les caractéristiques de la population à l'étude.⁸ Au lieu de cela, lorsque les chercheurs utilisent des estimations de la taille d'effet dans leurs calculs de la taille d'échantillon, ils

devraient plutôt utiliser des normes acceptées au sein des communautés de recherche afin d'identifier les valeurs représentatives et normatives attendues de la taille d'effet.¹⁴

Deuxièmement, comme cela a été mentionné plus haut, l'une des raisons pour lesquelles les chercheurs en éducation soulignent l'utilité de la taille d'effet dans le calcul de la taille d'échantillon est liée aux types de critères d'évaluation utilisés. La recherche en éducation implique généralement des mesures dépendantes qui ne peuvent être facilement observées, telles que les connaissances cliniques ou procédurales, l'empathie, les aptitudes de communication, etc. Plusieurs de ces résultats éducationnels sont souvent définis sur place par l'équipe de recherche et varient par conséquent d'une institution à une autre. C'est pour cette raison que les chercheurs proéminents en éducation médicale recommandent fortement aux chercheurs d'éviter de mettre au point de nouveaux outils de mesure spécifiques à leurs institutions, à moins que cela ne soit absolument nécessaire. Tout comme la recherche clinique nécessite une analyse rigoureuse afin de déterminer les critères d'évaluation les mieux adaptés en fonction des questions de recherche et de la méthodologie de l'étude, le même soin doit être apporté lorsqu'on tente d'identifier les critères adaptés pour les concepts éducationnels à l'étude. Lorsque cela est possible, il convient d'utiliser des instruments déjà validés dans la littérature. Dans les cas où les chercheurs ressentent le besoin de mettre au point un nouvel outil, cet outil doit non seulement être mis au point et calibré de façon précise, mais des données probantes de fiabilité et de validité doivent également être colligées avant de pouvoir utiliser cet outil dans une étude interventionnelle.⁶ Si un chercheur applique un outil de mesure inapproprié ou imprécis, il lui sera difficile de trouver un effet, peu importe la taille d'effet estimée.

Troisièmement, comme c'est le cas en recherche clinique, il est important de tenir compte du taux d'abandon potentiel lorsqu'on détermine la taille d'échantillon nécessaire pour une étude éducationnelle. De tels problèmes d'abandon sont souvent plus importants lorsqu'on retient une méthodologie d'étude longitudinale, c'est pourquoi il est important de bien planifier la taille d'échantillon dans de tels cas. Quatrièmement, il est important d'éviter d'amalgamer la magnitude d'une taille d'effet et l'importance réelle de l'effet. Bien que les interventions éducationnelles comportent souvent des tailles d'effet plus importantes que les interventions cliniques, cela ne veut pas dire qu'elles sont plus pertinentes. Pour paraphraser Wilkinson, de petits effets reproductibles avec des critères d'évaluation pertinents peuvent être particulièrement importants, alors que de

grands effets obtenus avec des critères triviaux pourraient être moins importants.¹⁵

Enfin, les estimations de la taille d'effet sont souvent associées à un certain degré d'incertitude, lequel est lié à plusieurs facteurs, notamment le bagage éducationnel des étudiants, les variations aléatoires de l'humeur et de l'attention, ou d'autres facteurs encore. Dès lors, le consensus au sein de la communauté de l'éducation et des sciences sociales veut que l'on rapporte, lorsque possible, les intervalles de confiance associés aux tailles d'effet.⁹ Les intervalles de confiance donnent une fourchette de valeurs possibles pour le paramètre à l'étude (dans ce cas, la taille d'effet) avec un certain degré de confiance. En rapportant les intervalles de confiance pour les tailles d'effet, on souligne non seulement la magnitude de l'effet, mais on communique également des informations quant à la précision de l'estimation. L'imprécision de l'estimation de la taille d'effet peut avoir un effet délétère sur le calcul de la taille d'échantillon.

Conclusion

Tout comme c'est le cas en recherche clinique, la taille d'échantillon est une composante importante de la conception et de la réalisation de recherches en éducation de qualité élevée. Dans les contextes d'éducation en soins de santé, l'utilisation de la taille d'effet dans le calcul de la taille d'échantillon est bien établie. Si l'on intègre rapidement la taille d'échantillon dans la conception de l'étude, on peut obtenir des données plus pertinentes qui auront en bout de ligne un impact plus important sur la pratique éducationnelle. Dans certains cas, des études multicentriques seront nécessaires, mais il vaut mieux déterminer cela tôt dans le processus de conception de l'étude afin de gagner du temps et d'éviter des frustrations plus tard.

Résolution de la mise en situation

Si l'on revient à la mise en situation présentée au début de cet article, vous aimeriez déterminer combien de participants seront nécessaires pour savoir si des examens hebdomadaires de divers formats (par ex., réponses courtes, questions à choix multiples) promeuvent la rétention des connaissances à long terme sur l'analgésie par bolus intermittent. Après avoir passé en revue la littérature, vous n'avez pas trouvé de mesures établies de la variabilité et des différences anticipées entre les deux groupes (par ex. examens répétés vs étude répétée). Toutefois, une méta-analyse récente portant sur des

interventions éducationnelles en éducation supérieure a révélé que les formats d'examens mixtes produisaient des tailles d'effet importantes (g de Hedges = 0,81).¹⁶ À l'aide de la version abrégée de la formule¹⁷ illustrée dans l'annexe :

$$n = 16 \frac{1}{\text{taille d'effet}^2} = 16 \frac{1}{0,81^2} = 16 \frac{1}{0,66} = 24,24$$

En résolvant la formule on obtient $n = 24,24$ ou 25 dans chaque groupe, ce qui donne une taille d'échantillon totale de 50 étudiants.

Conflicts of interest None declared.

Editorial responsibility This submission was handled by Dr. Hilary P. Grocott, Editor-in-Chief, *Canadian Journal of Anesthesia*.

Conflit d'intérêt Aucun.

Responsabilité éditoriale Cet article a été traité par Dr Hilary P. Grocott, rédacteur en chef, *Journal canadien d'anesthésie*.

Appendix Sample size and effect size formulae

Standard sample size formulae

There are several methods used to calculate sample size, depending on the type of data and the study design. The standard formula for comparing means between two groups, using a continuous dependent variable, is as follows:

$$n = 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})\sigma}{\delta} \right]^2$$

Here, δ represents the anticipated difference between group means (e.g., $\mu_1 - \mu_2$), and σ represents the pooled standard deviation between the two groups. $Z_{\alpha/2}$ is the value associated with the significance criterion, while Z_{β} is the value associated with statistical power of the study. For a type-I error of 0.05 (two-tailed), the corresponding $Z_{\alpha/2}$ is 1.96; and a statistical power of 0.8, the corresponding Z_{β} is 0.842. Note that this formula is the sample size required per group.

Simplified sample size formulae

Because the values associated with the significance criterion ($Z_{\alpha/2}$) and statistical power are conventionally set as 0.05 and 0.8, respectively, the formula above can be simplified to:

$$n = 16 \left(\frac{\sigma}{\delta} \right)^2$$

Where does the “16” come from? The full formula, as described by Lehr,¹⁷ is as follows:

$$\begin{aligned} n &= 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})\sigma}{\delta} \right]^2 = 2 \left[\frac{(1.96 + 0.842)\sigma}{\delta} \right]^2 \\ &= 2(1.96 + 0.842)^2 \left[\frac{\sigma}{\delta} \right]^2 = 15.70 \left[\frac{\sigma}{\delta} \right]^2 \approx 16 \left[\frac{\sigma}{\delta} \right]^2 \end{aligned}$$

Applying effect size estimates within sample size calculations

Some authors have argued (Norman *et al.*,¹⁸ Allen,¹⁹ Cohen¹¹) that in the absence of specific information about δ or σ , sample size calculations can be estimated using standardized effect sizes. For two-group comparisons, effect size is calculated as follows:

$$\text{effect size} = \frac{(\mu_1 - \mu_2)}{\sigma} = \frac{\delta}{\sigma}$$

where $\mu_1 - \mu_2$ is the difference between means (e.g., δ) and σ is the standard deviation. Therefore, effect size estimates can be used when calculating sample size for comparing group means. When using conventional values of $\alpha = 0.05$ and $\beta = 0.20$:

$$\begin{aligned} n &= 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})\sigma}{\delta} \right]^2 = 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})\sigma}{(\mu_1 - \mu_2)} \right]^2 \\ &= 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})}{\left(\frac{\mu_1 - \mu_2}{\sigma} \right)} \right]^2 = \frac{2(1.96 + 0.842)^2}{(\text{effect size})^2} \\ &= 16 \frac{1}{\text{effect size}^2} \end{aligned}$$

For example, for an effect size of 0.5, then $16/(0.5)^2 = 64$ subjects per group will be needed. Hence a total of 128 subjects are required. Using Cohen's¹¹ classification of effect sizes of 0.20, 0.50, and 0.80 as “small”, “medium”, and “large”, respectively. On this basis, the norm for sample sizes would be 400, 64, and 25 respectively, as evidenced by the following:

$$n = 16 \frac{1}{\text{effect size}^2} = 16 \left(\frac{1}{0.2^2} \right) = 16 \left(\frac{1}{0.04} \right) = 16(25) = 400$$

$$n = 16 \frac{1}{\text{effect size}^2} = 16 \left(\frac{1}{0.5^2} \right) = 16 \left(\frac{1}{0.25} \right) = 16(4) = 64$$

$$n = 16 \frac{1}{\text{effect size}^2} = 16 \left(\frac{1}{0.8^2} \right) = 16 \left(\frac{1}{0.64} \right) = 16(1.56) = 25$$

Annexe Formules pour la taille d'échantillon et la taille d'effet

Formule standard pour la taille d'échantillon

Il existe plusieurs méthodes pour calculer la taille d'échantillon, selon le type de données et la méthodologie de l'étude. La formule standard pour comparer des moyennes entre deux groupes, à l'aide d'une variable dépendante continue, est la suivante :

$$n = 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})\sigma}{\delta} \right]^2$$

Ici, δ représente la différence anticipée entre les moyennes de groupe (par ex., $\mu_1 - \mu_2$) et σ représente l'écart type pondéré entre les deux groupes. $Z_{\alpha/2}$ est la valeur associée au niveau de confiance, alors que Z_{β} est la valeur associée à la puissance statistique de l'étude. Pour une erreur de type I de 0,05 (bilatéral), le $Z_{\alpha/2}$ correspondant est 1,96; et une puissance statistique de 0,8, le Z_{β} correspondant est 0,842. À noter que cette formule donne la taille d'échantillon requise par groupe.

Formule simplifiée pour la taille d'échantillon

Parce que les valeurs associées au critère de portée ($Z_{\alpha/2}$) et à la puissance statistique sont traditionnellement établies à 0,05 et 0,8, respectivement, la formule ci-dessus peut être simplifiée comme suit :

$$n = 16 \left(\frac{\sigma}{\delta} \right)^2$$

D'où vient le chiffre « 16 »? La formule complète, telle que décrite par Lehr,¹⁷ est la suivante :

$$\begin{aligned} n &= 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})\sigma}{\delta} \right]^2 = 2 \left[\frac{(1,96 + 0,842)\sigma}{\delta} \right]^2 \\ &= 2(1,96 + 0,842)^2 \left[\frac{\sigma}{\delta} \right]^2 = 15,70 \left[\frac{\sigma}{\delta} \right]^2 \approx 16 \left[\frac{\sigma}{\delta} \right]^2 \end{aligned}$$

Application des estimations de la taille d'effet dans les calculs de la taille d'échantillon

Selon certains auteurs (Norman *et coll.*,¹⁸ Allen,¹⁹ Cohen,¹¹ en l'absence d'informations spécifiques concernant δ ou σ , les calculs de la taille d'échantillon peuvent être estimés à l'aide de tailles d'effet standardisées). Pour une comparaison entre deux groupes, la taille d'effet se calcule comme suit :

$$\text{taille d'effet} = \frac{(\mu_1 - \mu_2)}{\sigma} = \frac{\delta}{\sigma}$$

où $\mu_1 - \mu_2$ est la différence entre les moyennes (par ex., δ)

et σ est l'écart type. Par conséquent, les estimations de taille d'effet peuvent être utilisées lorsqu'on calcule la taille d'échantillon pour comparer des moyennes de groupe. Lorsqu'on utilise les valeurs conventionnelles de $\alpha = 0,05$ et $\beta = 0,20$:

$$\begin{aligned} n &= 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})\sigma}{\delta} \right]^2 = 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})\sigma}{(\mu_1 - \mu_2)} \right]^2 \\ &= 2 \left[\frac{(Z_{\alpha/2} + Z_{\beta})}{\left(\frac{\mu_1 - \mu_2}{\sigma} \right)} \right]^2 = \frac{2(1,96 + 0,842)^2}{(\text{taille d'effet})^2} \\ &= 16 \frac{1}{\text{taille d'effet}^2} \end{aligned}$$

Par exemple, pour une taille d'effet de 0,5, alors $16/(0,5)^2 = 64$ personnes par groupe sont nécessaires. On a donc besoin d'un total de 128 sujets. En utilisation la classification des tailles d'effet de Cohen,¹¹ où 0,20, 0,50 et 0,80 équivalent à « petite », « moyenne » et « grande », respectivement. Sur cette base, la norme pour des tailles d'échantillon serait de 400, 64 et 25, respectivement, comme le démontrent les formules suivantes :

$$n = 16 \frac{1}{\text{taille d'effet}^2} = 16 \left(\frac{1}{0,2^2} \right) = 16 \left(\frac{1}{0,04} \right) = 16(25) = 400$$

$$n = 16 \frac{1}{\text{taille d'effet}^2} = 16 \left(\frac{1}{0,5^2} \right) = 16 \left(\frac{1}{0,25} \right) = 16(4) = 64$$

$$n = 16 \frac{1}{\text{taille d'effet}^2} = 16 \left(\frac{1}{0,8^2} \right) = 16 \left(\frac{1}{0,64} \right) = 16(1,56) = 25$$

References

1. Larsen DP, Butler AC. Test-enhanced learning. In: Walsh K (Ed.). Oxford Textbook of Medical Education. Oxford University Press; 2013: 443-52.
2. Harden RM, Grant J, Buckley G, Hart IR. BEME guide no. 1: Best Evidence Medical Education. Med Teach 1999; 21: 553-62.
3. Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant 2010; 25: 1388-93.
4. Stansfield RB, Gruppen L. Power analyses: planning, conducting and evaluating education research. In: Cleland J, Durning SJ, editors. Researching Medical Education. Oxford: Wiley Blackwell; 2015. p. 43-8.
5. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10: 407-15.
6. Blanchard RD, Artino AR Jr, Visintainer PF. Applying clinical research skills to conduct education research: important recommendations for success. J Grad Med Educ 2014; 6: 619-22.
7. Cohen J. Statistical Power Analysis for the Behavioral Sciences. NY: Academic Press; 1977.
8. Sullivan GM, Feinn R. Using effect size—or why the p value is not enough. J Grad Med Educ 2012; 4: 279-82.
9. Kelley K, Rausch JR. Sample size planning for the standardized mean difference: accuracy in parameter estimation via narrow confidence intervals. Psychol Methods 2006; 11: 363-85.

10. *Hattie J.* The applicability of visible learning to higher education. *Scholarsh Teach Learn Psychol* 2015; 1: 79-91.
11. *Cohen J.* Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988 .
12. *Erdfelder E, Faul F, Buchner A.* GPOWER: a general power analysis program. *Behav Res Methods Instrum Comput* 1996; 28: 1-11.
13. *Zheng J, Li Y, Lin T, Estrada A, Lu X, Feng C.* Sample size calculations for comparing groups with continuous outcomes. *Shanghai Arch Psychiatry* 2017; 29: 250-6.
14. *Lenth RV.* Some practical guidelines for effective sample size determination. *Am Stat* 2001; 55: 187-93.
15. *Wilkinson L.* Statistical methods in psychology: guidelines and explanations. *Am Psychol* 1999; 54: 594-604.
16. *Adesope OO, Trevisan DA, Sundararajan N.* Rethining the use of tests: a meta-analysis of practice testing. *Rev Educ Res* 2017; 87: 659-701.
17. *Lehr R.* Sixteen S-squared over D-squared: a relation for crude sample size estimates. *Stat Med* 1992; 11: 1099-102.
18. *Norman G, Monteiro S, Salama S.* Sample size calculations: should the emperor's clothes be off the peg or made to measure? *BMJ* 2012; 345: e5278.
19. *Allen JC.* Sample size calculations for two independent groups: a useful rule of thumb. *Statistics. Proceedings of Singapore Healthcare* 2011; 20: 138-40.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.